Penalized projection estimators of the Aalen multiplicative intensity.

Patricia REYNAUD-BOURET
Georgia Institute of Technology
E-mail: Patricia.Reynaud-Bouret@ens.fr

6th December 2002

Abstract

We study the problem of non parametric estimation of the intensity of counting processes satisfying the Aalen multiplicative intensity model. To do so, we use model selection techniques and more precisely penalized projection estimators for a random inner product. For histogram estimators, under some assumptions on the process, we obtain adaptive results for the minimax risk. In general, for more intricate (predictable) models, we only obtain oracle inequalities. The study is completed by some simulations in the right-censoring model.

AMS Classification: 62G07, 62M09.
Keywords: penalized projection estimators, model selection, counting processes, multiplicative intensity model.

1 Introduction

Let us first present the processes with Aalen multiplicative intensity. Let $(N_t)_{t\geq 0}$ be a counting process i.e. a nondecreasing random piecewise constant function with $N_0 = 0$. This process generates classically a filtration $(\mathcal{F}_t)_{t\geq 0}$ and with respect to this filtration, $(N_t)_{t\geq 0}$ has a compensator $(\Lambda_t)_{t\geq 0}$, i.e. a nondecreasing random function such that $(M_t = N_t - \Lambda_t)_{t\geq 0}$ is a martingale.

The counting process $(N_t)_{t\geq 0}$ verifies the Aalen multiplicative intensity model if we can write:

$$d\Lambda_t = Y_t s(t) dt, \tag{1.1}$$

where $Y = (Y_t)_{t\geq 0}$ is a nonnegative predictable process and where s is a deterministic function.

The purpose of this paper is to estimate the intensity s on $[0, \tau]$ using the observations of $(N_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$. Let us first give some examples of processes with Aalen multiplicative intensity.

The easiest case is the time Poisson process. It corresponds to the case where Y is constant. Many works deal with the non parametric estimation of the intensity of a Poisson process: let us mention the work of M. Rudemo [22] for histogram and kernel estimators, the work of W.-C. Kim and J.-Y. Koo [12] for wavelet estimators and the work of L. Cavalier and J.-Y. Koo [9] for thresholding procedures. Model selection procedures are also used in [19].

Another simple example is the process defined by $(N_t = \mathbb{1}_{X \leq t}, t \geq 0)$ where X is a positive random variable with density f. This process has only one jump and verifies (1.1) with $Y_t = \mathbb{1}_{X \geq t}$ and with $s(t) = f(t)/\mathbb{P}(X \geq t)$. The function s is the **hazard rate** of X. If X represents the life time of some patient, $s(t)$ represents the probability that the patient stays alive after time t if he is alive at time t.

The observations of life times can sometimes be censored. This is the case when the patient goes away from the hospital study. Then the death time is not observed but we know that he was alive when leaving the study. This situation is modeled by some other positive random variable U independent of X and the observations are the variables $T = X \wedge U$ and $D = \mathbb{1}_{T=X}$. This model is known as the **right-censoring model** with independent censorship. Then the process $N_t = D\mathbb{1}_{T \leq t}$ has an Aalen multiplicative intensity (1.1) with $Y_t = \mathbb{1}_{T \geq t}$ and with s the hazard rate of X.

If we dispose of a n-sample of counting processes, $N^1, ..., N^n$ with corresponding predictable processes $Y^1, ..., Y^n$ and with same intensity s, we can study the **aggregated process** N with predictable process Y, where both are defined by

$$N_t = \sum_{i=1}^{n} N_t^i \text{ and } Y_t = \sum_{i=1}^{n} Y_t^i \text{ for all } t \geq 0. \tag{1.2}$$

For instance in the right-censoring model, the aggregated Y is a non increasing process with integer values and with $Y_0 = n$ the number of observations.Let us fix t, then Y_t represents the number of events (deaths or departures) which happen after time t.

The problem of estimating the hazard rate in this model is well known. For instance, let us mention the work of A. Antoniadis, G. Grégoire and G. Nason [3] for wavelets estimators on sieves and the references therein and the work of C. Kooperberg, C.J. Stone and Y.K. Truong [13] for splines estimators. Moreover there exist some model selection adaptive procedures due to S. Döhler and L. Rüschendorf [10]. The same method is used by G. Castellan and F. Letué [8] for the Cox model with right-censorship.

Many other examples of processes with multiplicative intensity are mentioned in the book of P.K. Andersen, O. Borgan, R. Gill and N. Keiding [1]. For instance, if $(X_t)_{t \geq 0}$ is a Markov process with finite state space, the counting process $(N_t^{hj}, t \geq 0)$, where N_t^{hj} represents the number of transitions from h to j before time t, has a multiplicative intensity of the form (1.1) where s is the transition intensity from h to j and where Y is defined by $(Y_t = \mathbb{1}_{X(t)=h}, t \geq 0)$. We can also dispose of a n-sample of i.i.d. counting processes corresponding to each individual Markov processes and aggregate them by (1.2). The aggregated process Y is again integer valued and bounded by n: at time t, Y_t represents the number of individuals in state h. This situation models for instance the healthy-diseased transitions (see Example I.3.10 in [1]).

There exist also cases where the process cannot be divided into individual processes, and so cannot be written as in (1.2). This is the case for the model of Drosophila flies matings proposed in Example III.1.10 of [1]. However this model verifies the multiplicative intensity property (1.1) with a Y, which, in some sense, always corresponds to a number of events which may arrive after time t and which is bounded by a known constant.

Many papers consider the problem of estimating s in the general Aalen multiplicative intensity model. One of the first paper on this subject is due to H. Ramlau-Hansen [18] who proves consistency and asymptotic normality results for some kernel estimators with fixed bandwidth. G. Grégoire [11] gives a data-driven criterion to choose the bandwidth of the Ramlau-Hansen estimators by cross-validation and proves also consistency and asymptotic

normality results. Other possible estimators are maximum likelihood estimators on some sieve whose rate of convergence were studied by S. Van de Geer [23]. A. Antoniadis [2] chooses the sieve by penalization and proves consistency and asymptotic normality for penalized maximum likelihood estimators, the penalization depending on the regularity of the functions.

In this paper we want to go further in this direction by constructing estimators of the intensity using the observations of N_t and Y_t with as few assumptions as possible on the process and on the intensity. In particular, we do not want to assume a precise knowledge of the regularity of the function. Moreover we want to give non asymptotic results, since in practice the observations come generally from medical survey and their numbers is very small.

By scale changes, the problem can be reduced to the estimation of s on $[0,1]$ and to the observation of the processes on $[0,1]$.

This work actually gives an extension of the results of S. Döhler and L. Rüschendorf in [10] to the general Aalen multiplicative intensity model. This also proves the adaptivity of the resulting estimators for some particular cases.

The method is inspired by the work of L. Birgé and P. Massart [6] on penalized projection estimators (**p.p.e.**) for density estimation.

Before describing the method, let us first present a classical assumption on the process:

Assumption 1. *Y is bounded by a known A.*

For instance, in the right-censoring model or in the Markovian model, $A = n$ represents the number of observations.

Now let us define the least-square contrast in our context: for all f in $\mathbb{L}^2([0,1], dt)$,

$$\gamma_A(f) = -2\int_0^1 f(t)\frac{dN_t}{A} + \int_0^1 f^2(t)Y'_t dt, \tag{1.3}$$

where $Y'_t = Y_t/A$. This is not the contrast used by S. Döhler and L. Rüschendorf [10]: they use a log-likelihood contrast which is much more intricate to deal with than the least-square one, though it is celebrated and gives good results.

The projection estimator of s on a finite dimensional linear subspace S is then defined by

$$\hat{s} = \mathrm{argmin}_{f\in S}\gamma_A(f). \tag{1.4}$$

The associated random norm is defined for all f in $\mathbb{L}^2([0,1], dt)$ by

$$\|f\|^2_{\mathrm{rand}} = \int_0^1 f^2(t)Y'_t dt. \tag{1.5}$$

If $\{h_\lambda, \lambda \in \Gamma\}$ is an orthonormal basis of S for this random norm, we can simplify the projection estimator and write:

$$\hat{s} = \sum_{\lambda\in\Gamma}\left(\int_0^1 h_\lambda(t)\frac{dN_t}{A}\right)h_\lambda. \tag{1.6}$$

If we want to estimate s by a projection estimator, we have to correctly choose the subspace S. If we want to do some adaptive estimation, this subspace or **model** must be chosen via a data driven criterion. To achieve this goal, we introduce a family of models (finite dimensional linear subspaces) $\{S_m, m \in \mathcal{M}_A\}$ and we associate to each S_m the

projection estimator $\hat{s}_m$ of s on it. Let us take a penalty denoted pen, which is a positive function on $\mathcal{M}_A$, independent of s and, if necessary, random. Then to find a good model, it is sufficient to minimize the following data driven criterion:

$$\hat{m} = \text{argmin}_{m \in \mathcal{M}_A} \left(\gamma_A(\hat{s}_m) + \text{pen}(m) \right), \tag{1.7}$$

and the **penalized projection estimator (p.p.e.)**, $\tilde{s}$ is defined by $\hat{s}_{\hat{m}}$.

A good penalty is a penalty which gives **oracle type inequality**, i.e an inequality of the form:

$$\|s - \tilde{s}\|^2 \leq C \inf_{m \in \mathcal{M}_A} \left(\|s - s_m\|^2 + \text{pen}(m) \right), \tag{1.8}$$

for some positive constant C where s_m is the projection of s on S_m for $\|.\|$, where this inequality holds either in probability or in expectation with, if necessary, the addition of some negligible term.

The norm $\|.\|$ in (1.8) can be either the random norm $\|.\|_{\text{rand}}$ or the deterministic norm defined for all f in $\mathbb{L}^2([0,1], dt)$ by:

$$\|f\|^2_{\text{det}} = \int_0^1 f^2(t) \mathbb{E}(Y'_t) dt. \tag{1.9}$$

If $\|s - s_m\|^2 + \text{pen}(m) \simeq \|s - \hat{s}_m\|$, we say that one has a true oracle inequality. This means that up to some constant the p.p.e. $\tilde{s}$ does as well as the best possible estimator in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$ without knowing s. This proves the adaptivity of the p.p.e. in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$.

The choice of the family of models is also very important. For the density estimation [6], the models used have to be such that the infinity norm of the functions f in the model S_m is well controlled by their $\mathbb{L}^2$-norm. Here we encounter the same problem and we need the same control between the infinity norm and the random norm. Therefore we have restricted ourselves to the cases where we know an orthonormal basis of the models for the random norm. Then we can easily check on this basis that the coefficients in the orthonormal decomposition (and consequently the $\mathbb{L}^2$-norm) have good comparison vis a vis the infinity norm of the functions. We deal thus with two cases.

The first one is the histogram case. The basis is clear but we must restrict ourselves to aggregated processes (1.2) to be able to control the variance term of the estimator. This is done in Section 2. Under these assumptions, we are able to prove oracle type inequality for well chosen penalties and families of models. We are also able to prove minimax results.

For other deterministic models, there is apparently no clear comparison between the random norm and the infinity norm in the model.

The second case deals with random predictable models. They are built as follows. If $\{\varphi_\lambda, \lambda \in m\}$ is a classical deterministic orthonormal basis of $\mathbb{L}^2([0,1], dt)$, typically a part of a Fourier basis, then $\{h_\lambda = \varphi_\lambda / \sqrt{Y'}, \lambda \in m\}$ becomes, when Y' is positive, an orthonormal basis for the random product. The model $S_m = \text{Span}\{h_\lambda, \lambda \in m\}$ is therefore a random predictable subspace where we can ensure good comparison between infinity norm and random norm if the $\{\varphi_\lambda, \lambda \in m\}$ are well chosen (Fourier or wavelet basis). One of the advantages of these models is to allow us to remove the aggregation assumption among many other technical assumptions. We prove oracle type inequality for this case in Section 3.

Section 4 contains simulation studies of those two strategies in the right-censoring model and compares these results with already existing estimators.

The last part of the present paper is devoted to the proofs of the main results.

2 Histogram quasi-least square estimators

The purpose of this section is to treat deterministic histogram models. We therefore make the following assumptions:

Assumption 2.

- *N is an aggregated process (see (1.2)), with predictable process Y and with individual processes $N^1, ..., N^n$ and $Y^1, ..., Y^n$.*
- *Each Y^i is bounded by* 1.
- *The number of individual jumps of the N^i is bounded by a known positive constant K.*

For instance the right-censoring model verifies these assumptions with $K = 1$ but the Markovian models do not verify the last of these assumptions.

Under these assumptions, A defined by Assumption 1 is equal to n. We will also assume that there exists an unknown constant R such that s is bounded by R.

If the Y^i's are just bounded by a known B, it is sufficient to divide the Y^i's by B and to estimate sB to verify Assumption 2.

We can compare these assumptions with those of G. Grégoire [11]. He assumes N to be aggregated and Y to be bounded too. He does not assume a bound on N. Thus he can manage Markovian models but he assumes that n/Y is bounded by a quantity independent of n. This is not needed here.

2.1 Study on one model

Under Assumption 2, the least-square contrast becomes

$$\gamma_A(f) = -\frac{2}{n}\int_0^1 f(t)dN_t + \frac{1}{n}\int_0^1 f^2(t)Y'_t dt.$$

Let us now construct the projection histogram estimator. Let m be a partition of $[0, 1]$. For all I in m, let

$$b_I = (1/n)\int_0^1 \mathbb{1}_I Y_t dt.$$

Then for the random norm $\|.\|_{\text{rand}}$, the family $\{\mathbb{1}_I/\sqrt{b_I}, I \in m\}$ is an orthonormal basis of the subspace of piecewise constant functions on m. Let $|m|$ be the number of intervals in the partition m.

We notice that b_I depends only on the observations. Let $\beta_I = \mathbb{E}(b_I)$ and let N_I be the number of points of N lying in I.

Let $\mathcal{I}_m$ be the set of intervals I in m such that the b_I's are larger than $1/n^2$.
The **quasi-least square** histogram estimator on S_m, the space of the piecewise constant functions on m, is the projection estimator of s defined by (1.4) on S'_m the set of piecewise constant functions on m, null outside $\mathcal{I}_m$. Using (1.6) this estimator can be rewritten as:

$$\hat{s}_m = \sum_{I\in\mathcal{I}_m} \frac{N_I}{nb_I}\mathbb{1}_I. \tag{2.1}$$

It is more convenient to deal with this quasi-least square estimator (i.e. the projection estimator of s on S'_m) than with the projection estimator of s on S_m, because $\hat{s}_m$ is bounded.

Risk of the quasi-least square estimator:

Let s_m be the projection of s on S_m for the random scalar product:

$$s_m = \sum_{I \in m} \frac{a_I}{b_I} \mathbb{1}_I,$$

where

$$a_I = (1/n) \int_0^1 \mathbb{1}_I Y_t s(t) dt.$$

Note that if $b_I = 0$ then $a_I = 0$ and the corresponding coefficient of s_m is zero. Let $\alpha_I = \mathbb{E}(a_I)$ and let s'_m be the projection of s on S'_m:

$$s'_m = \sum_{I \in \mathcal{I}_m} \frac{a_I}{b_I} \mathbb{1}_I.$$

Finally we denote by s_m^{det} the projection of s on S_m for the deterministic scalar product:

$$s_m^{\text{det}} = \sum_{I \in m} \frac{\alpha_I}{\beta_I} \mathbb{1}_I.$$

The distance between s and $\hat{s}_m$ can be divided in the following way:

$$\|s - \hat{s}_m\|^2_{\text{rand}} = \|s - s'_m\|^2_{\text{rand}} + \|s'_m - \hat{s}_m\|^2_{\text{rand}}. \tag{2.2}$$

The first term is a bias term which, unlike for density estimation (see [6]), is now random. We can bound it by:

$$\begin{aligned} \|s - s'_m\|^2_{\text{rand}} &= \|s - s_m\|^2_{\text{rand}} + \|s_m - s'_m\|^2_{\text{rand}} \\ &\leq \inf_{t \in S_m} \|s - t\|^2_{\text{rand}} + \frac{R^2}{n}, \end{aligned} \tag{2.3}$$

assuming that there is less intervals than n i.e. $|m| \leq n$. Therefore the expectation of the bias term is less than

$$\mathbb{E}(\|s - s'_m\|^2_{\text{rand}}) \leq \inf_{s \in S_m} \mathbb{E}(\|s - t\|^2_{\text{rand}}) + \frac{R^2}{n} = \|s - s_m^{\text{det}}\|^2_{\text{det}} + \frac{R^2}{n},$$

which decreases when the intervals of the partition become small.

The behavior is very different for the second term in (2.2) whose expectation is called the variance term. For a set of intervals $\mathcal{T}$, let us set:

$$\chi^2_{\mathcal{T}} = \sum_{I \in \mathcal{T}} \frac{(\frac{N_I}{n} - a_I)^2}{b_I}. \tag{2.4}$$

Then the second term in (2.2) is exactly $\chi^2_{\mathcal{I}_m}$. If we assume that the b_I's are close to their expectation denoted by β_I, and assumed to be non zero, $\chi^2_{\mathcal{I}_m}$, χ^2_m and

$$Z^2_m = \sum_{I \in m} \frac{(\frac{N_I}{n} - a_I)^2}{\beta_I}, \tag{2.5}$$

are also really close. Let us also assume that Z^2_m is close to its expectation which is

$$\mathbb{E}(Z^2_m) = \sum_{I \in m} \frac{\alpha_I}{n \beta_I}.$$

Then this expectation lies between $r|m|/n$ and $R|m|/n$, if s is upper bounded by R and lower bounded by r.

If all the previous approximations are licit, the variance term must grow like the dimension of the model S_m when the bias term decreases.

2.2 Penalized least square histograms

If we want to find a good model, we must balance the bias term and the variance term but we must do this through a data driven criterion, without knowing s.

Let $\{S'_m, m \in \mathcal{M}_A\}$ be the family of models corresponding to the family of partitions $\mathcal{M}_A$ of $[0,1]$.

The best partition or the best model, the one which we would choose if we knew s is called the **oracle** and is defined by:

$$\begin{aligned} \bar{m} &= \operatorname{argmin}_{m\in\mathcal{M}_A} \mathbb{E}(\|s-\hat{s}_m\|^2_{\text{rand}}) \\ &= \operatorname{argmin}_{m\in\mathcal{M}_A} \mathbb{E}(-\|s'_m\|^2_{\text{rand}} + \|s'_m-\hat{s}_m\|^2_{\text{rand}}) \\ &\simeq \operatorname{argmin}_{m\in\mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|^2_{\text{rand}} + 2\|s'_m-\hat{s}_m\|^2_{\text{rand}}) \\ &\simeq \operatorname{argmin}_{m\in\mathcal{M}_A} \mathbb{E}(-\|\hat{s}_m\|^2_{\text{rand}} + 2\chi^2_{\mathcal{I}_m}). \end{aligned} \tag{2.6}$$

The symbol $\simeq$ indicates that the expectations are not equal but that if the coefficients of $\hat{s}_m - s'_m$ are close to zero, the expectations are close to each other.

Moreover we have,

$$\|\hat{s}_m\|^2_{\text{rand}} = -\gamma_A(\hat{s}_m).$$

Hence in estimating the previous quantities, we are going to choose the model $\hat{m}$ given by (1.7) with the penalty that verifies "pen(m) is an estimate of two times the variance term".

The equation (2.6) corresponds to the ISE minimization for kernel estimators done by G. Grégoire who does an unbias estimation of the risk by a leave-one out procedure.

Here we choose the partition by the general penalized data-driven criterion given in (1.7). But to prove that a penalty is well chosen we have to prove an inequality of the type (1.8). Hence we need to understand how far away $\hat{m}$ can be from the oracle. More precisely, we have to understand the behavior of $\chi^2_{\mathcal{I}_m}$ and see if the penalty overestimate it or not.

2.3 Control of the chi-square statistic

The behavior of the $\chi^2_{\mathcal{I}_m}$'s are however very difficult to control. Thus we bound them by

$$\chi^2_{\mathcal{I}_m} \leq Z_m^2 V_m, \text{ for all } m \in \mathcal{M}_A,$$

where Z_m^2 is given by (2.5) and where $V_m = \sup_{I\in m} \frac{\beta_I}{b_I}$.

Moreover, the square root of Z_m^2 can be seen to be

$$Z_m = \sup_{\sum_{I\in m}\delta_I^2\beta_I=1} \left(\frac{1}{n}\sum_{i=1}^n \int_0^1 \left(\sum_{I\in m} \delta_I \mathbb{1}_I(t) \right) (dN_t^i - Y_t^i s(t)dt) \right).$$

We can thus apply the recent version of Talagrand's inequality obtained by E. Rio [21].

Proposition 1. *Under Assumption 2, for all $\varepsilon, x > 0$,*

$$\mathbb{P}\left(Z_m \geq (1+\varepsilon)\sqrt{\sum_{I\in m}\frac{\alpha_I}{n\beta_I}} + \sqrt{2v_m\frac{x}{n}} + (1/2+\varepsilon^{-1})b\frac{K+R}{n}x \right) \leq e^{-x},$$

where

$$b = \sup_{I \in m} \frac{1}{\sqrt{\beta_I}}, \qquad R \geq \|s\|_\infty,$$

and where

$$v_m = \sup_{\delta / \sum_{I \in m} \delta_I^2 \beta_I = 1} \int_0^1 \sum_{I \in m} \delta_I^2 \mathbb{1}_I(t) \mathbb{E}(Y_t^1) s(t) dt,$$

is bounded by

$$R_m = \sup_{I \in m} \frac{\alpha_I}{\beta_I}.$$

Proof. Applying Theorem 1.4 of [17] to

$$''X''_{i,\delta} = \frac{1}{n} \int_0^1 (\sum_{I \in m} \delta_I \mathbb{1}_I(t))(dN_t^i - Y_t^i s(t) dt)$$

which are centered variables, it is very easy to derive the previous bound, knowing that the number of jumps of the N^i is bounded by K and that Y^i is bounded by 1.
We can restrict the supremum to a countable dense family of δ in order to carefully apply the result of Rio. But by density, this leads to the present result. ■

We can also find a large event on which the behavior of Z_m is sub-Gaussian, as P. Massart does for estimating the density of an i.i.d. n-sample in [15].

Proposition 2. *Let ε be a positive number and let $\Omega_m(\varepsilon)$ be the event*

$$\Omega_m(\varepsilon) = \left\{ \forall I \in m, |(N_I/n) - a_I| \leq \left(\frac{2\varepsilon}{(K+R)(1/2 + \varepsilon^{-1})} \right) \beta_I \right\}.$$

Then under Assumption 2, for all positive x,

$$\mathbb{P}\left(Z_m \mathbb{1}_{\Omega_m(\varepsilon)} \leq (1+\varepsilon) \left(\sqrt{\sum_{I \in m} \frac{\alpha_I}{n \beta_I}} + \sqrt{\frac{2 R_m x}{n}} \right) \right) \leq e^{-x},$$

where

$$R_m = \sup_{I \in m} \frac{\alpha_I}{\beta_I}.$$

Proof. We know that Z_m is attained at $\hat{\delta}$ such that for all I,

$$\hat{\delta}_I = \frac{(N_I/n) - a_I}{\beta_I Z_m}.$$

Hence on $\Omega_m(\varepsilon) \cap \{Z_m \geq z\}$,

$$Z_m = \sup_{\substack{\delta / \sum_{I \in m} \delta_I^2 \beta_I = 1, \\ \sup_{I \in m} \delta_I \leq \frac{2\varepsilon}{(K+R)(1/2+\varepsilon^{-1})z}}} \frac{1}{n} \int_0^1 (\sum_{I \in m} \delta_I \mathbb{1}_I(t))(dN_t^i - Y_t^i s(t) dt).$$

If we apply Talagrand's inequality to this last supremum with $z = \sqrt{(2R_m x)/n}$, we get precisely the previous result. ■

We can obtain the same kind of result replacing R_m by every upper bound on R_m.

The last part of Assumption 2 is a technical assumption precisely here to derive Propositions 1 and 2. If we had a Bernstein's type inequality instead of Talagrand's inequality for the suprema, we might be able to remove the assumption of the existence of K and treat also Markovian models too by the same method.

2.4 Oracle inequalities

We can now construct oracle inequalities. The first one is a bound in probability on a large event, for the random norm. The second one is an expectation bound for the deterministic norm.

Theorem 1. *Let N be a counting process with multiplicative intensity $Y_t s(t)$ (see (1.1)) satisfying Assumption 2. Assume that s is bounded by an unknown positive R.*
Let Γ be a fixed regular partition of $[0,1]$ (i.e. constructed on equally spaced points). Let $\mathcal{M}_A$ be a family of partitions which are constructed with unions of intervals of Γ. For a given penalty pen *on $\mathcal{M}_A$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.4)). Assume that:*

1. *there exist μ and ρ strictly positive such that $\inf_{I\in\Gamma}(|\Gamma|\alpha_I) \geq \mu$ and $\inf_{I\in\Gamma}(|\Gamma|\beta_I) \geq \rho$,*
2. *there exists a finite family of positive weights on $\mathcal{M}_A$, $(L_m)_{m\in\mathcal{M}_A}$ such that $\sum_{m\in\mathcal{M}_A} \exp(-L_m|m|) \leq \Sigma$, for some Σ independent of n,*
3. *$|\Gamma|$ is less than $n/\ln^2 n$.*

Let $d > 1$. One sets for all m in $\mathcal{M}_A$,

$$\text{pen}(m) = d\tilde{R}_\Gamma \frac{|m|}{n}\left(1+\sqrt{2L_m}\right)^2$$

where

$$\tilde{R}_\Gamma = \sup_{I\in\Gamma}\frac{N_I}{nb_I}.$$

Then there exists a large event $\Omega(d)$ such that for all η positive, there exists positive continuous functions C, C' and C'' such that

$$\mathbb{P}(\Omega(d)^c) \leq C''(d,K,R,\rho,\mu)/n^\eta$$

and such that on $\Omega(d)$, for all $\xi > 0$ with probability larger than $1-\Sigma e^{-\xi}$,

$$\|s-\tilde{s}\|^2_{\text{rand}} \leq C(d)\inf_{m\in\mathcal{M}_A}\left\{\|s-s'_m\|^2_{\text{rand}} + \frac{|m|L_m}{n}R_\Gamma\right\} + C'(d)R_\Gamma\frac{\xi}{n},$$

where $R_\Gamma = \sup_{I\in\Gamma}\frac{\alpha_I}{\beta_I}$.

Corollary 1. *Under the previous assumptions and notations, for the penalized least square histogram estimator described in Theorem 1 there exist positive continuous functions H and H' such that*

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\text{det}}) \leq H(d)\inf_{m\in\mathcal{M}_A}\left(\|s-s_m^{\text{det}}\|^2_{\text{det}} + R_\Gamma\frac{|m|L_m}{n}\right) + \frac{H(d,R,K,\rho,\mu,\Sigma)}{n}.$$

The weights L_m can be constant if the family of partitions has for instance at most one model per dimension. Then these oracle type inequalities become true oracle inequalities and the p.p.e. is adaptive in the family $\{\hat{s}_m, m\in\mathcal{M}_A\}$.

The oracle type inequality of Theorem 1 is a probability bound. It is therefore a stronger result than the one of Corollary 1. But for the minimax risk, it is better to have oracle inequality for deterministic loss function (here $\|.\|^2_{\text{det}}$).

We can compare this penalized model selection to the model selection built in [10] for the right-censoring case. In [10], the penalty is very large (in $\exp(\exp(R))$) and depends on the knowledge of a bound on s. Here the penalty is linear in R and as we deal with histogram estimators, we can estimate the bound on s by $\tilde{R}_\Gamma$. We see in the simulations (see Section 4), that when the penalty is too large, the estimator behaves poorly and $C(d)$ becomes very large. However the estimators built in [10] apply to various types of models while we can only prove oracle inequalities for histogram estimators.

The weights L_m are here to take into account the complexity of the family of models. We refer to [7] for an extensive list of applications of these weights. Such applications could also be done here, but we cannot see their performances since we restrict ourselves to the histogram models.

2.5 Minimax risk

The oracle inequalities imply that the p.p.e. is adaptive in this family $\{\hat{s}_m, m \in \mathcal{M}_A\}$: without knowing s it finds the best possible estimator in the family, up to some multiplicative constant for the risk. But we may also want to compare it with all the other possible estimators. This is the aim of this minimax study.

We know that the histograms have good approximation properties for α-hölderian functions with $0 < \alpha < 1$. Hence we hope that the p.p.e. given in Theorem 1 will have good minimax properties for such set of functions.
Let L and r be positive constants and let

$$\mathcal{H}_{L,\alpha,r} = \left\{ f \in L^2([0,1], dt) \ : \ \forall x, y \in [0,1], |f(x) - f(y)| \leq L|x-y|^\alpha \text{ and } r + L \geq f(x) \geq r \right\}.$$

Let the minimax risk on $\mathcal{H}_{L,\alpha,r}$ be defined by

$$R(\mathcal{H}_{L,\alpha,r}) = \inf_{\hat{s}} \sup_{s \in \mathcal{H}_{L,\alpha,r}} \mathbb{E}_s(\|s - \hat{s}\|^2_{\text{det}}),$$

where $\hat{s}$ describes all the possible estimators in $\mathbb{L}^2([0,1], dt)$. The minimax risk on $\mathcal{H}_{L,\alpha,r}$ represents the risk of the best estimator for the worst s to estimate in the family $\mathcal{H}_{L,\alpha,r}$.

Proposition 3. *If there exist μ and M such that for all s in $\mathcal{H}_{L,\alpha,r}$, $\mu \leq \mathbb{E}_s(Y_t^1) \leq M$, then there exists a positive continuous function c such that*

$$R(\mathcal{H}_{L,\alpha,r}) \geq c(\alpha) n^{-\frac{2\alpha}{2\alpha+1}} L^{\frac{2}{2\alpha+1}} r^{\frac{2\alpha}{2\alpha+1}} \mu M^{-\frac{2\alpha}{2\alpha+1}}.$$

The above assumptions are true in many situations. For instance, in the right-censoring model, $\mathbb{E}_s(Y_t^1)$ is less than 1 and larger than $\exp -(r+L)$.

We also remark that the exponent in n is the rate of convergence of the classical regression problem.

We now want to compare the risk of $\tilde{s}$ built in Theorem 1 with the minimax risk. Let us look at the following classical strategy: $|\Gamma| = 2^J$ is of order $n/\ln^2 n$ and we take the sub-partitions, m, of Γ which are also regular with 2^j intervals and j less than J. There is at most one model by dimension, so we can take constant weights ($L_m = 1$, for instance) to build the penalty and consequently the p.p.e. We call this strategy the **nested histogram strategy**. Now let us apply Corollary 1.

If s is in $\mathcal{H}_{L,\alpha,r}$, the bias $\|s - s_m^{\text{det}}\|_{\text{det}}^2$ is bounded by $L^2|m|^{-2\alpha}\varsigma$ where $\varsigma = \int_0^1 \mathbb{E}(Y_t^1)dt$. When n tends to infinity, we obtain taking m such that $|m|$ is of order $(n\varsigma L^2/R)^{1/(2\alpha+1)}$ (which is less than $|\Gamma|$ for n large enough)

$$\mathbb{E}(\|s - \tilde{s}\|_{\text{det}}^2) = O\left(n^{-\frac{2\alpha}{2\alpha+1}} L^{\frac{2}{2\alpha+1}} R^{\frac{2\alpha}{2\alpha+1}} \varsigma^{\frac{1}{2\alpha+1}}\right).$$

We can compare this bound to the lower bound found in Proposition 3. One gets the exact power in n and in L. The bound R on s replaces r the infimum of s. The quantity ς replaces $\mu^{2\alpha+1}M^{-2\alpha}$ and represents the order of magnitude of $\mathbb{E}(Y_t^1)$.

This means that $\tilde{s}$ without knowing α and L (depending on s) does as well as the best possible estimator which knows this fact. In this sense, $\tilde{s}$ is an **adaptive estimator** for the α-hölderian functions with $0 < \alpha < 1$.

3 Predictable models

We have seen what can easily be done for aggregated processes. Let us remove this assumption and deal with predictable models. We keep the notations introduced in (1.3) and (1.5).

We assume Assumption 1 and the following fact.

Assumption 3. *There exists c positive such that if $Y_t < c$, for some $t > 0$, then $Y_t = 0$.*

For a Poisson process, one has $A = c$. For the other examples, Y is an integer-valued function and $c = 1$ works.

The aggregated case leads us to think that A plays the same role as n. Consequently, the asymptotic point of view in this framework is "A tends to infinity". On the other hand, c is considered as a fixed constant, independent of A.

3.1 Construction and risk for one model

Let J_t be $\mathbb{1}_{Y_t \neq 0}$. The family of models is then built as follows: let $\{\varphi_\lambda, \lambda \in \Gamma\}$ be a classical orthonormal basis for $\mathbb{L}^2([0,1], dt)$; let $\mathcal{M}_A$ be a family of subsets of Γ. Then for m in $\mathcal{M}_A$, we set $S_m = \text{Span}\{h_\lambda(.) = \left(\varphi_\lambda(.)/\sqrt{Y_.'}\right) J_., \lambda \in m\}$. Let $\hat{s}_m$ be the projection estimator associated to S_m and defined by (1.4). Let also $|m|$ be the cardinality of m.

Let us also define the following observable event:

$$\Omega = \{\forall t \geq 0, Y_t \neq 0\}. \tag{3.1}$$

We will see later that in many situations, Ω has a very large probability to happen when A is large enough.

On Ω, the h_λ's form an orthonormal basis of S_m for the random scalar product and consequently $\hat{s}_m$ is of the form (1.6).

Risk of the projection estimator:

On Ω, the projection s_m of s on S_m for the random inner product is given by

$$s_m(.) = \sum_{\lambda \in m} \left(\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y_t'} dt \right) \frac{\varphi_\lambda(.)}{\sqrt{Y_.'}}.$$

Then we can write

$$\|s - \hat{s}_m\|_{\text{rand}}^2 = \|s - s_m\|_{\text{rand}}^2 + \|s_m - \hat{s}_m\|_{\text{rand}}^2.$$

The first term corresponds to a bias term and the expectation of the second term is a variance term.

The bias term is random as in the histogram case. We can write

$$\|s - s_m\|^2_{\text{rand}} = \int_0^1 \left(s(t)\sqrt{Y'_t} - \sum_{\lambda \in m} \left(\int_0^1 \varphi_\lambda(t) s(t) \sqrt{Y'_t} dt \right) \varphi_\lambda(t) \right)^2 dt.$$

Thus, the bias term corresponds to the classical $\mathbb{L}^2([0,1], dt)$ error when one projects $s\sqrt{Y'}$ on $\text{Span}\{\varphi_\lambda, \lambda \in m\}$. If m grows, this term generally decreases.

The second term corresponds to a χ^2-type statistics and behaves quite differently: on Ω, it is $\chi(m)^2_1$ where the process $(\chi(m)^2_t)_{t \geq 0}$ is defined by

$$\chi(m)^2_t = \sum_{\lambda \in m} \left(\int_0^t \frac{\varphi_\lambda(u)}{\sqrt{Y'_u}} J_u \frac{dM_u}{A} \right)^2, \text{ for all } t \geq 0. \tag{3.2}$$

Its compensator $(C(m)_t)_{t \geq 0}$ is defined by

$$C(m)_t = \sum_{\lambda \in m} \int_0^t \varphi^2_\lambda(u) s(u) J_u \frac{du}{A}, \text{ for all } t \geq 0.$$

But, on Ω, $C(m)_1$ is constant and moreover if $r \leq s \leq R$, then $\frac{r|m|}{A} \leq C(m)_1 \leq \frac{R|m|}{A}$. Hence, if $\chi(m)^2_1$ is close to $C(m)_1$, it increases like the dimension of the model.

3.2 Penalized projection estimator

Again, if we want to find a good model, we must balance the bias term and the variance term, but we must do this through a data driven criterion, without knowing s. Therefore we use (1.7) and we obtain $\tilde{s}$, the p.p.e. for the family of models $\{S_m, m \in \mathcal{M}_A\}$.

Here there also exists an heuristic argument. We can also defined an oracle, the best model, the one which we could choose if we knew s:

$$\begin{aligned} \bar{m} &= \text{argmin}_{m \in \mathcal{M}_A} \|s - \hat{s}_m\|^2_{\text{rand}} \\ &= \text{argmin}_{m \in \mathcal{M}_A} - \|s_m\|^2_{\text{rand}} + \|s_m - \hat{s}_m\|^2_{\text{rand}} \\ &\simeq \text{argmin}_{m \in \mathcal{M}_A} - \|\hat{s}_m\|^2_{\text{rand}} + 2\|s_m - \hat{s}_m\|^2_{\text{rand}} \\ &\simeq \text{argmin}_{m \in \mathcal{M}_A} - \|\hat{s}_m\|^2_{\text{rand}} + 2\chi(m)^2_1. \end{aligned} \tag{3.3}$$

The approximations ($\simeq$) are good if (as in the histogram case) the coefficients of $s_m - \hat{s}_m$ are close to their expectation which is 0. If $\chi(m)^2_1$ is close to $C(m)_1$, a penalty of the form $2c|m|/A$ would be convenient (where c is of the order of s). Again we found the factor 2 which always appears when doing this kind of heuristic and which is due to Mallows [14] in the Gaussian framework.

The study of the probabilistic behavior of $\chi(m)_t$ around its compensator has been made in [20].

3.3 Oracle inequalities

We can now derive oracle type inequalities for predictable models.

Theorem 2. *Let N be a counting process with multiplicative intensity $Y_t s(t)$ (see (1.1)) satisfying Assumptions 1 and 3.*

Let $\{S_m, m \in \mathcal{M}_A\}$ be a family of predictable models built as previously from the deterministic classical orthonormal family $\{\varphi_\lambda, \lambda \in \Gamma\}$. For a given penalty pen *on $\mathcal{M}_A$, let $\tilde{s}$ be the associated penalized projection estimator (see (1.7)).*
Assume that:

1. *there exists a positive constant Φ, such that for all m in $\mathcal{M}_A$,*
$$\| \sum_{\lambda \in m} \varphi_\lambda^2 \|_\infty \leq \Phi |m|.$$
2. *There exists a finite family of positive weights on $\mathcal{M}_A$, $(L_m)_{m \in \mathcal{M}_A}$ such that* $\sum_{m \in \mathcal{M}_A} |m|^2 \exp(-L_m) \leq \Sigma$,

Moreover assume that we know a bound on s denoted by R.

Let d be larger than 1. One sets for all m in $\mathcal{M}_A$:
$$\mathrm{pen}(m) = d \frac{|m|}{A} \left(\sqrt{R}(1 + 3\sqrt{2L_m}) + \sqrt{\frac{\Phi}{c}} L_m \right)^2 .$$

Then there exist positive continuous functions C and C' such that on Ω defined by (3.1), one has
$$\mathbb{E}(\|s - \tilde{s}\|_{\mathrm{rand}}^2 \mathbb{1}_\Omega) \leq C(d) \inf_{m \in \mathcal{M}_A} \left(\mathbb{E}(\|s - s_m\|_{\mathrm{rand}}^2) + \mathrm{pen}(m) \right) + \frac{C'(d, R, \Phi, c, \Sigma)}{A}.$$

As the models are random, we can only derive oracle inequalities for the random norm. Probability bounds exist but are much more intricate than in Theorem 1 (see Part 5).

The classical case is when $\{\varphi_\lambda, \lambda \in \Gamma\}$ is a Fourier basis $\{\exp(-2ik\pi x), k \in \mathbb{Z}\}$ with $\mathcal{M}_A = \{m_k = \{-k, k\}, k \geq 0\}$, then $|m_k| = 2k + 1$ and $L_{m_k} = 4 \ln k$. The constant Φ in the theorem is then equal to 1. In practice we must take a finite family of models i.e. take $k \leq A$ for instance.

We can also consider a wavelet basis $\{\varphi_{j,k}, j \geq 0, k \geq 0\}$ with regularity h and $\mathcal{M}_A = \{m_l, l \geq 0\}$ where $m_j = \{(l, k), l \leq j\}$. If the wavelet has finite support, Φ defined in Theorem 2 depends only on the choice of the basis.

As the family of models is nested in both previous cases, the penalty is of order $|m| R \log(|m|)/A$. Thus we recover an oracle inequality up to a logarithmic factor, since the variance term is of order $|m|/A$. We can think of more complex family of models (i.e. more models with the same dimension). If the number of models with dimension D in the family is of order a power of D, we can have the same kind of penalty and we also recover oracle inequalities up to a logarithmic factor. If the number of models with same dimension D is of order e^D, the penalty must be of order $R|m|^\gamma/A$, for $\gamma > 1$. It is really larger than the variance term and there is no longer an oracle inequality.

When one has an oracle inequality, one can also say that the p.p.e. is adaptive in the family $\{\hat{s}_m, m \in \mathcal{M}_A\}$. But as we do not know the approximation properties of the random spaces S_m, we cannot, in general, consider adaptivity properties in the minimax sense.

However, if N is a Poisson process, Y_t is a deterministic constant. This fact implies that all the norms are deterministic. In this case, let us assume that s belongs to
$$\mathcal{B}(\rho, L, B_{2,2}^\alpha) = \left\{ t = \rho + u \ : \ t \geq 0, \ \int_0^1 u \, dx = 0, \ u \in B_{2,2}^\alpha, \ \|u\|_{2,2}^\alpha \leq L \right\},$$

where $B^{\alpha}_{2,2}$ is the classical Besov space with regularity α, $1/2 \leq \alpha \leq h$ and with $\mathbb{L}^2$-norm. Consider the last strategy with a wavelet family of regularity h. Then compromising between the penalty and the bias in the oracle inequality, we obtain when A tends to infinity:

$$\mathbb{E}(\|s - \tilde{s}\|^2) = O\left(L^{\frac{2}{2\alpha+1}} R^{\frac{2\alpha}{2\alpha+1}} \left(\frac{A}{\ln^2 A}\right)^{-\frac{2\alpha}{2\alpha+1}}\right).$$

This is the minimax rate (see [19]) up to the logarithmic factor and the replacement of $\int_0^1 s$ by R. Therefore, the resulting p.p.e. is adaptive in the minimax sense for all the Besov balls with regularity less than h, up to a logarithmic factor.

This logarithmic factor is actually not needed in the Poisson case: in [19], it is proved that penalties of the type $R|m|/A$ with the same previous families of models gives oracle inequalities without logarithmic factor and consequently minimax rate without logarithmic factor. If we apply Theorem 2, which is valid for more general processes, the weights L_m are constant and the last term explodes with Σ for large families of models: there is no longer an oracle inequality.

The same kind of remark can be made if we want to use a more complex family of models (i.e. more models with the same dimension in the family of models). In the Poisson framework, there exist penalties of the type $R|m|(\log A)/A$ which are proved to give up to some logarithmic factor oracle inequalities. Applying Theorem 2 to the same type of strategies give an explosive last term. However, the counting processes are very well adapted to biomedical data. In such cases, the number of observations $n \simeq A$ is not very large and if we also take a small number of models, there is no longer an explosive phenomenon. This justifies the interest in having non-asymptotic results.

3.4 Improvement

Estimation of R: The fact that the penalty depends on the knowledge of a bound on s can be a nuisance. In some cases, we can estimate this bound.

Let Γ be a regular partition of $[0,1]$. Suppose that s is L, α-hölderian, and let s_Γ be the projection of s for the random norm on the space of histograms with partition Γ. Then

$$\|s - s_\Gamma\|_\infty \leq L|\Gamma|^{-\alpha}.$$

Take $|\Gamma|$ of order $A/\ln^2 A$. Then $\|s\|_\infty \leq \|s_\Gamma\| + o(1)$, when A goes to infinity. But

$$\|s_\Gamma\|_\infty = \sup_{I \in \Gamma} \frac{\int_I s(t) Y'_t dt}{\int_I Y'_t dt}.$$

So we can replace R by $(1+\varepsilon)\tilde{R}_\Gamma$ where

$$\tilde{R}_\Gamma = \sup_{I \in \Gamma} \frac{N_I}{A \int_I Y'_t dt},$$

if we are on

$$\Omega(\varepsilon) = \left\{ \left| \int_I dM_t/A \right| \leq \frac{\varepsilon}{1+\varepsilon} \int_I s(t) Y'_t dt \right\}.$$

The complementary of this last event is very small (it has probability of order $o(A^{-\eta})$, for all $\eta > 0$) if we assume the process to be aggregated and Assumption 2 (or moment

assumptions). Then we can apply Bernstein's inequality to $\int_I dM/A$ and to $\int_I s(t)Y'_t dt$. On $\Omega \cap \Omega(d)^c$, the estimator is bounded and one can conclude as in the proof of Corollary 1.

Magnitude of Ω: In the aggregated cases, Ω is a very large event and we can also give an oracle type inequality for $\mathbb{E}(\|s - \tilde{s}\|^2_{\text{rand}})$.

Let us look more precisely at the right-censoring model. In this case $A = n$ and $Y'_t = \sum_{i=1}^n \mathbb{1}_{X_i \wedge U_i \geq t}$, where the X_i's are the life times and the U_i's form the censorship. It can be seen as $1 - \hat{F}_n(t)$, where $\hat{F}_n(t)$ is the empirical cumulative distribution function associated to the $X_i \wedge U_i$'s. One has

$$\forall \lambda > 0, \mathbb{P}\left(\sqrt{n} \sup_{t \in \mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \geq \lambda\right) \leq 2e^{-2\lambda^2},$$

where F is the true cumulative distribution function of the $X_i \wedge U_i$'s (see [16]).

Thus if we assume that there exists a positive μ such that $\mathbb{E}(Y^1_t) \geq \mu > 0$ on $[0,1]$, then

$$\Omega^c \subset \{\sup_{t \in \mathbb{R}} \left|Y'_t - \mathbb{E}(Y^1_t)\right| \geq \mu/2\},$$

and $\mathbb{P}(\Omega^c) \leq 2\exp(-n\mu^2/2)$.

Hence we can defined the estimators on the whole probability space by:

$$\hat{s}_m(.) = \sum_{\lambda \in m} \left(\int_0^1 \frac{\varphi_\lambda(t)}{\sqrt{Y'_t}} J_t \frac{dN_t}{A}\right) \frac{\varphi_\lambda(.)}{\sqrt{Y'_.}} J_.,$$

and this even if we are not in Ω. This estimator is a projection estimator only on Ω. We do the model selection as in Theorem 2. As these estimators are always bounded, we proceed as in Corollary 1 and we can bound $\mathbb{E}(\|s - \tilde{s}\|^2_{\text{rand}})$ (on the whole probability space) by the same kind of bound as in Theorem 2.

4 Simulations

The aim of this section is not to provide an extensive simulation study but to just present an illustration of the previous methods. We restrict ourself to the right-censoring model (see the Introduction).

The life times $X_1, ..., X_n$ are generated for a given hazard rate s on $[0,1]$. The censorship $U_1, ...U_n$ are generated as uniform variables on $[0,2]$. We observe $T_i = X_i \wedge U_i$ and $D_i = \mathbb{1}_{T_i = X_i}$, for all i less than n. Some of the T's will be outside of $[0,1]$: this is a good case since it ensures that we are on the event Ω of the previous section.

At the end of the interval, there are sometimes few points to see, so it is well known that all the estimators of the hazard rate become inefficient. To take this into account, the random norm $\|s - \tilde{s}\|^2_{\text{rand}}$ is a good quantity since the multiplication by Y' gives less weight to the end of the interval. Moreover this norm is not just convenient, it is very close to the Kullback-Leibler distance as we have seen when we have done minimax computations in Section 2.5 and in Section 5. This random norm is denoted by "Risk" on the figures.

Five different methods are compared here and this only for some examples. We do not pretend to give precise formula for the penalty but just some penalties which work quite well.

The **regular histogram strategy** (R.H.S.) consists in taking all the regular partitions of $[0,1]$ up to a certain number of intervals which is the minimum of the number of

observations and of 20. The factor 20 is present to have small computing times. As there is one model by dimension and as there exist a big partition Γ (even if we do not know precisely its form) such that all these partitions are sub-partitions of Γ, Theorem 1 allows us to have constant weights. A penalty of the following form is therefore convenient

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n},$$

where $\tilde{R} = \sup_{I\in\Lambda} N_I/(nb_I)$, with Λ which is just the thinest partition of our family of models replacing Γ in order to simplify the computations. This estimator $\tilde{R}$ of the supremum of s is used for all the other strategies in the penalty.

The **exhaustive histogram strategy** (E.H.S.) looks at all the sub-partitions of Γ where Γ is a regular partition with d intervals where d is the minimum of 8 and of the integer part of $[n/\log(n)^2]$. Again, the factor 8 is present to have small computing times. The penalty is of the form

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n}\left(1+\sqrt{2\log\left(\frac{n}{|m|}\right)}\right)^2,$$

i.e. the weights L_m of Theorem 1 are of the form $\log(n/|m|)$ to ensure the convergence of Σ.

The **progressive histogram strategy** (P.H.S.) is specially created to take into account the fact that we have a poor estimation near 1. It consists in taking partitions whose intervals are small near 0 and large near 1. More precisely, we add to the family of regular partitions i.e. the partitions of the form $\{0, 1/N, 2/N, ..., (N-1)/N, 1\}$, partitions which progress in a polynomial way, i.e. of the form $\{0, 1^k/N^k, 2^k/N^k, ..., (N-1)^k/N^k, 1\}$, and also partitions which progress in an exponential way, i.e of the form $\{0, k^1/k^N, k^2/k^N, ..., k^{N-1}/k^N, 1\}$. We take this for all the integers k less than 3 and all the integers N less than the minimum of 20 and of the number of observations divided by 3. Once more, the factors 3 and 20 give us small computing times. As for the (R.H.S), Theorem 1 allows us to have constant weights, and we therefore use a penalty of the following form

$$\text{pen}(m) = \frac{d\tilde{R}|m|}{n}.$$

The **Fourier strategy** (F.S.) is the strategy described in the previous section. The φ_λ's form the Fourier basis and we consider the nested models described in this part. According to Theorem 2, the penalty is of the form

$$\text{pen}(m) = \frac{d|m|}{n}(\sqrt{\tilde{R}_\Gamma} + \log|m|)^2.$$

In order to have simple formulas, we also delete the second term of the penalty which is smaller than the other terms.

If d is well chosen in all the previous strategies, the penalized criteria must estimate the risks of each projection estimator and be close to those risks up to the additive term $\|s\|^2_{\text{rand}}$. The last strategy is then the **minimal criteria strategy** (M.C.S.) which chooses between the four previous estimators the one with the smallest penalized criteria.

Of course before computing this last strategy (M.C.S), we have to find good parameters d for R.H.S, E.H.S, P.H.S and F.S. which ensure that the penalized criteria are close to the risks of the projection estimators.

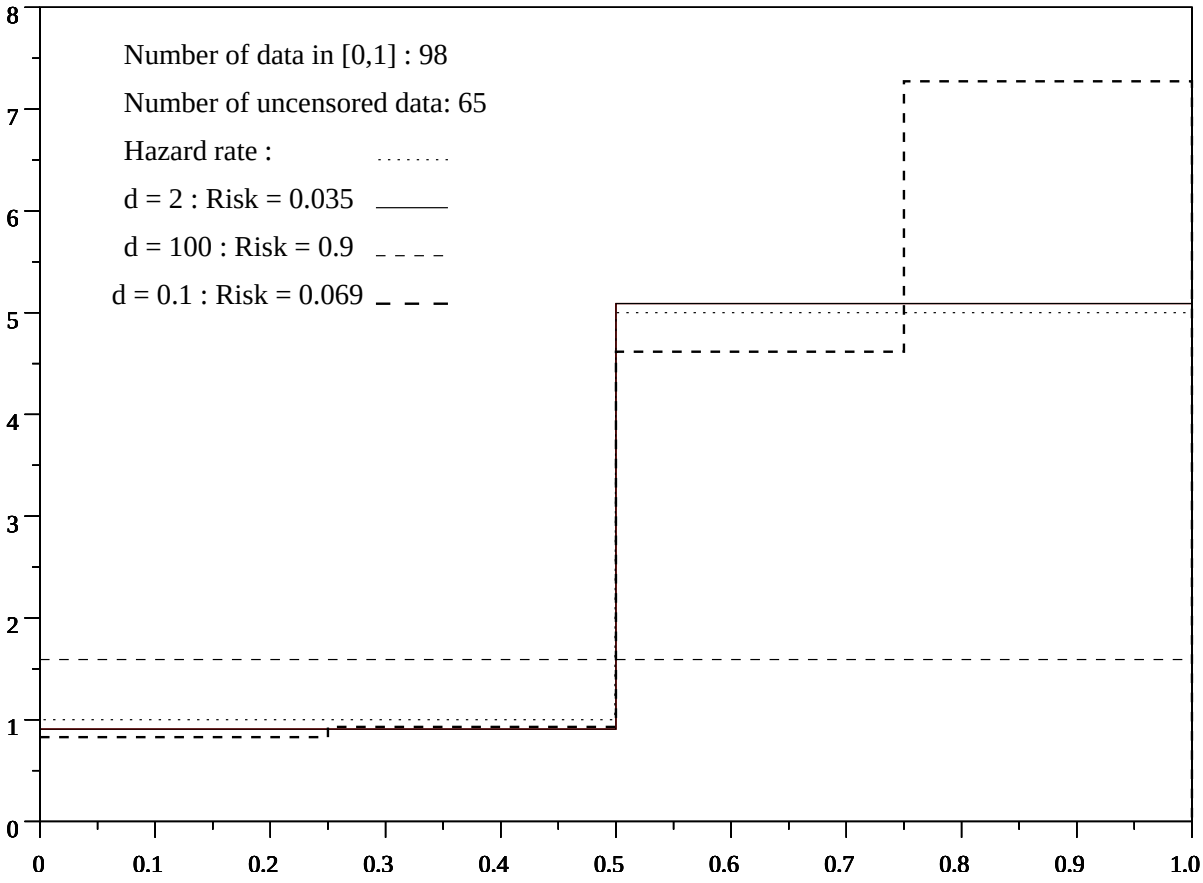

Figure 1: Example of the influence of d for the regular histogram strategy

4.1 Influence of d

To illustrate the influence of d, let us look at the easiest strategy: the regular histogram strategy (R.H.S.).

In Figure 1, the unknown hazard rate is already an histogram. If the penalty is equal to $2\tilde{R}|m|/n$ (i.e. $d = 2$), the p.p.e. recovers the right model with two intervals. If $d = 100$, the procedure gives a very large data-driven criteria for large dimension and the p.p.e. finds the model with only one interval. If $d = 0.1$, the penalty is not large enough. The p.p.e. finds a model with too many intervals (here 4). The best case (or the oracle, i.e. the model chosen if we knew s) is the one found by the procedure for $d = 2$.

For the classical model selection techniques in Gaussian frameworks [7], we might have very poor estimators for $d < 1$, and quite good estimators for $d > 1$. But here this dichotomy is not so precise. Actually there exists a large interval of possible d around 1 which work well. This is probably due to the presence of $\tilde{R}$ which could overestimate the size of s and which is not present in the penalty in the Gaussian case of [7] since, there, even the variance term does not depend on the function to estimate.

However a Mallows type heuristics (i.e. with $d = 2$) seems to work very well for the R.H.S. on a lot of examples.

There exists methods to estimate a proper d by a data-driven criterion in the Gaussian framework. This is the work of E. Lebarbier in her PhD Thesis. We do not try here to do

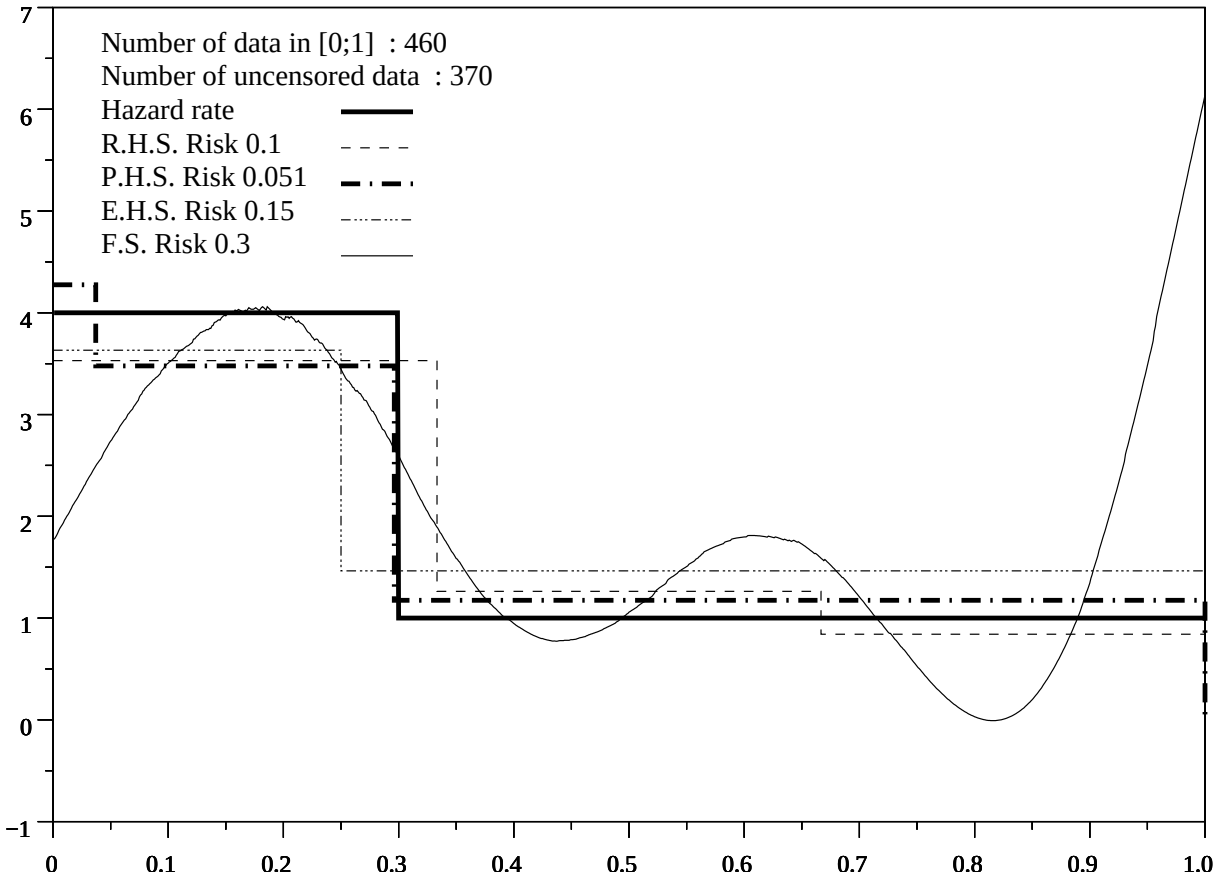

Figure 2: Results for a piecewise constant function. (M.C.S.=P.H.S)

this kind of work which is long and complicated.

For the E.H.S, the P.H.S and the F.S., we have found good d's with the same kind of analysis as before. In these cases, there also exist intervals of possible d on which the estimator is close to the oracle. However for the E.H.S and for the F.S., $d = 2$ is too large. This is probably due to the fact that the penalty contains a logarithmic term which cannot be removed: if we try to remove it, the p.p.e. would be under-penalized and would find models with too large dimension.

For the P.H.S, $d = 2$ is too small. This is probably due to the fact that there is much more than one model by dimension even if we can also take constant weights L_m.

Hence in all the following simulations we set:

- for the R.H.S. $d = 2$.
- for the E.H.S. $d = 0.4$.
- for the P.H.S. $d = 2.5$.
- for the F.S. $d = 1$.

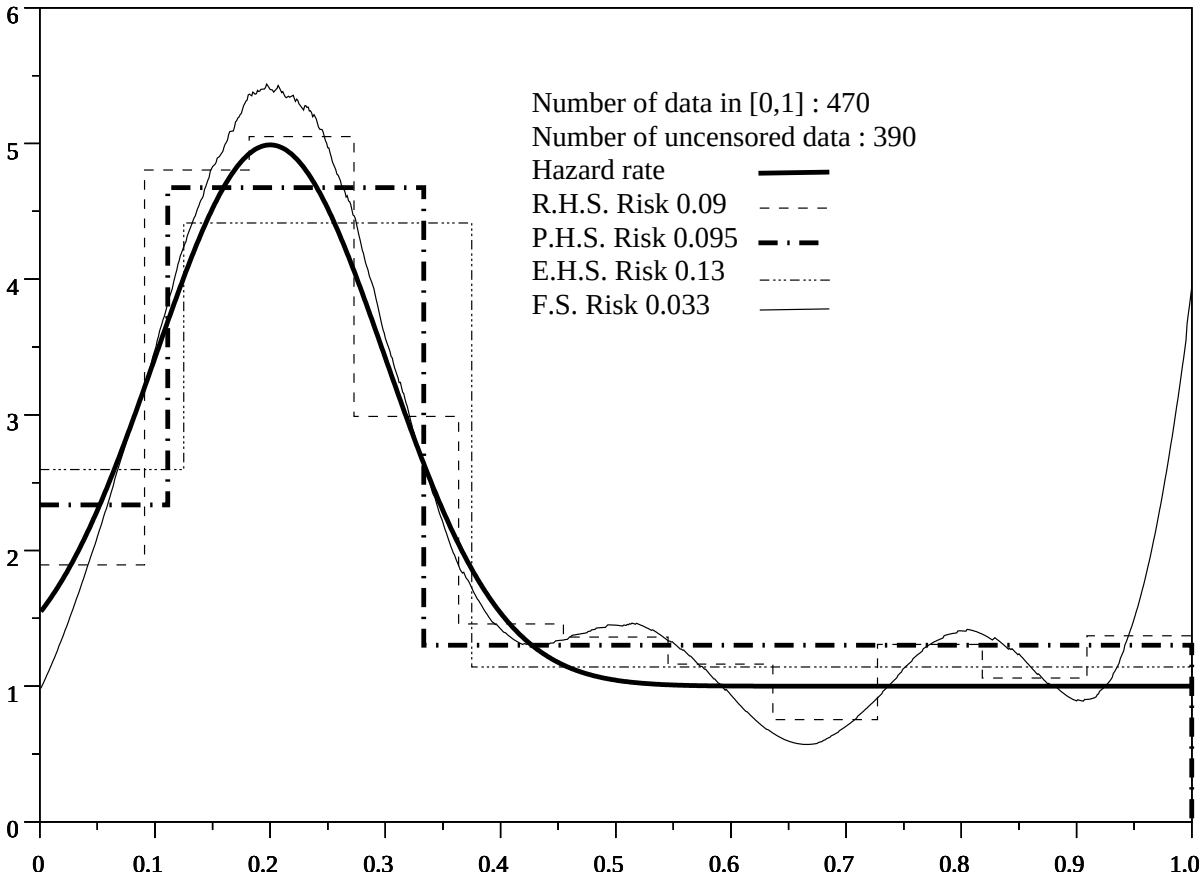

Figure 3: Results for a smooth function. (M.C.S.=F.S)

4.2 Comparison of the different strategies

We have computed the risk for the four methods on various sets of functions. Figures 2 and 3 present what happens for two particular examples where we clearly see the differences between the three kind of estimators by histograms. The M.C.S. method gives in both case the estimator with minimum risk (i.e. the P.H.S. for Figure 2 and the F.S. for Figure 3).

Figures 4 and 5 present the hazard rate functions and Figure 6 gives the risk of the different estimators. More precisely, for each kind of hazard rates with a censorship which is uniform on $[0, 2]$, we simulate either 200 or 500 data. All the results are given in mean over 200 accumulations.

Some of the simulated observations are bigger than 1 and this is the reason why we indicate the number of data which are strictly between 0 and 1 and also the number of uncensored data. We also give the most frequent choice of the M.C.S. to see if it corresponds to the minimum of the risk. Of course the risk of the M.C.S. is not exactly the one of the most frequent choice because sometimes the M.C.S. chooses something different. When two strategies are chosen with approximately the same frequency by the M.C.S., we have given both names.

The first remark is that, for a fixed hazard and a fixed method, the risk seems to decrease with the number of variables. Moreover this risk is proportional to s as we can see for the functions 1 and 2. This has to be taken into account when comparing the results for different types of functions.

The risk is larger when the function is not in the family of models we are using, and this happens even when the function is piecewise constants but its partition does not belong to the family of partitions we have taken. For instance it explains why for the function 4 the risk of the E.H.S is bigger for 500 observations than for 200: the way we have built the biggest partition Γ gives a regular partition with 8 intervals for 500 observations and with 7 intervals for 200 observations. In this last partition there is one point close to 0.3, which no longer exists when we take 500 observations.

In general, for piecewise constant hazard rates, the histograms family are better, and this is the choice of the M.C.S. When the function is smoother, the F.S. seems sometimes better, and this is also the choice of the M.C.S.

The P.H.S. seems to work well even for smooth function. This is probably because it is very well adapted to see differences in behavior near the origin: for instance it detects more easily the bump in function 11 (see also function 13) than the F.S. which is not localized and has a tendency to move when the hazard rate is flat (see also Figure 3).

The E.H.S. does not seem so useful at a first look but in some cases (Function 6, 500 observations, for instance), it is in fact similar to the P.H.S. and as therefore the same risk. In terms of criteria, as there is a logarithm factor for the penalty of the E.H.S., the M.C.S. will always prefer the P.H.S. Another explanation for the fact that the P.H.S. seems better than the E.H.S. is that we cannot trust any estimator at the end of the interval. The P.H.S. which already takes this fact into account and simplifies its family of models gives better estimation. In other words, the E.H.S. has a tendency to look for precision at the end of the interval, but it is not useful since what it finds at the end of the interval, should not be really trusted. This phenomenon is due to the fact that we have right censored data; for other type of counting processes, this would probably not happen.

Globally the M.C.S. gives a good way to decide between all the strategies: even if it

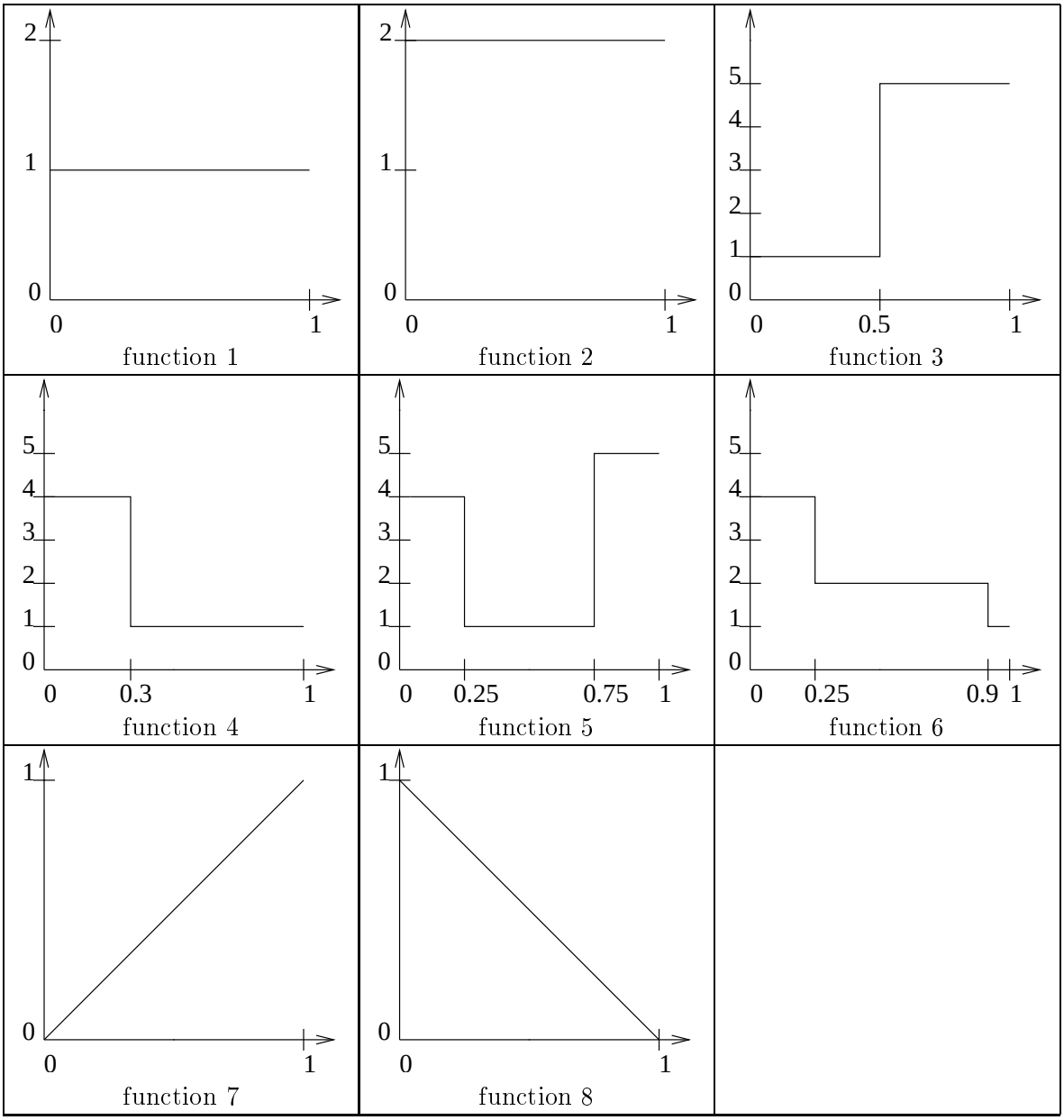

Figure 4: Hazard rates to estimate (1st part).

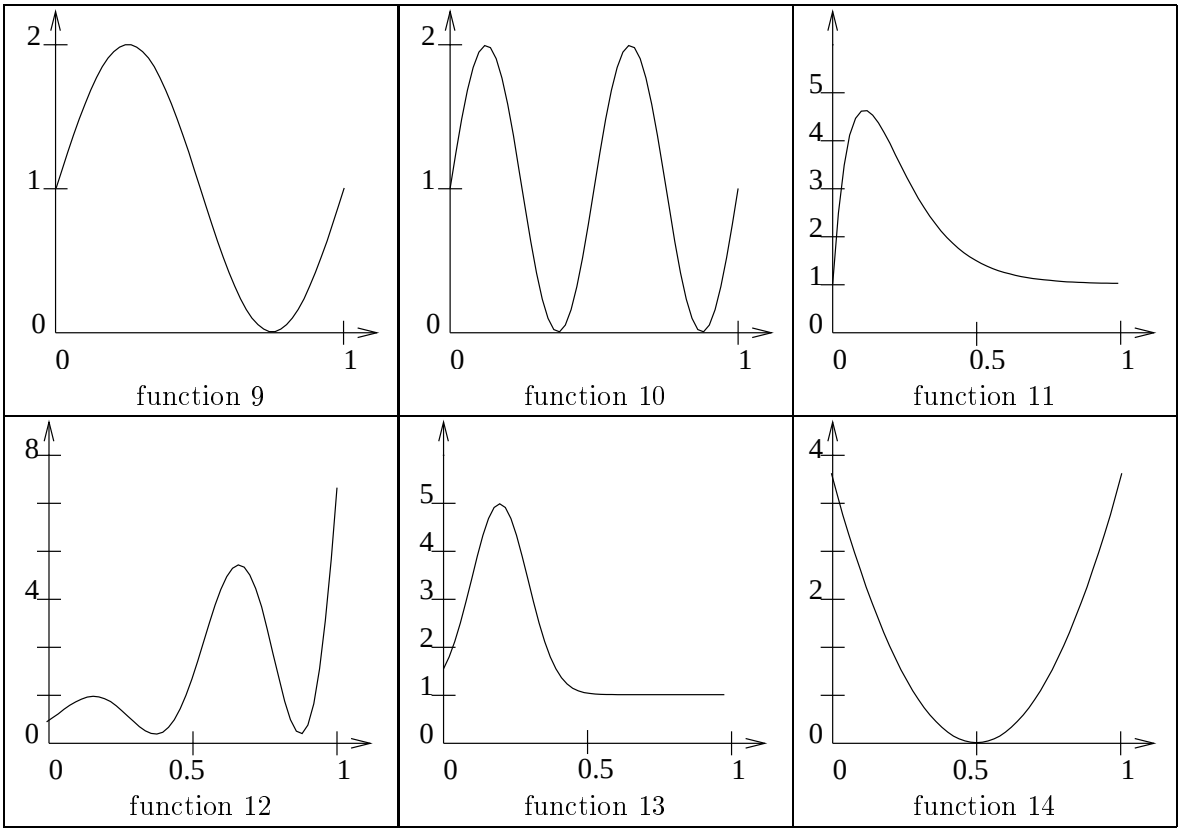

Figure 5: Hazard rates to estimate (2nd part).

Hazard rates	Data in $[0,1]$		Uncensored data		R.H.S.		P.H.S.		E.H.S.		F.S.		M.C.S.		Choice of
	200	500	200	500	200	500	200	500	200	500	200	500	200	500	M.C.S.
1	163	408	99	249	0.007	0.003	0.008	0.004	0.004	0.002	0.032	0.020	0.014	0.005	R.H.S.
2	186	467	142	357	0.013	0.005	0.015	0.006	0.010	0.004	0.119	0.072	0.023	0.008	R.H.S.
3	195	487	144	360	0.03	0.01	0.03	0.01	0.4	0.01	0.28	0.16	0.04	0.01	R.H.S.
4	185	462	152	380	0.107	0.08	0.09	0.04	0.06	0.13	0.31	0.21	0.09	0.04	P.H.S.
5	193	484	159	400	0.07	0.02	0.13	0.02	0.43	0.02	0.40	0.23	0.12	0.02	R.H.S.
6	190	477	160	402	0.10	0.03	0.07	0.02	0.14	0.02	0.33	0.22	0.07	0.02	P.H.S.
7	140	350	54	134	0.02	0.01	0.02	0.01	0.04	0.02	0.03	0.01	0.03	0.01	F.S./R.H.S.
8	140	349	66	168	0.02	0.01	0.02	0.01	0.03	0.01	0.04	0.03	0.02	0.01	R.H.S.
9	163	407	108	271	0.06	0.04	0.06	0.04	0.06	0.05	0.03	0.02	0.04	0.03	F.S.
10	163	408	104	261	0.06	0.05	0.09	0.05	0.20	0.05	0.03	0.02	0.04	0.02	F.S.
11	188	469	157	391	0.22	0.17	0.17	0.10	0.23	0.19	0.21	0.10	0.18	0.10	P.H.S.
12	186	465	131	327	0.25	0.17	0.26	0.19	0.34	0.21	0.10	0.07	0.12	0.07	F.S.
13	186	465	152	380	0.26	0.12	0.16	0.11	0.27	0.15	0.10	0.06	0.14	0.07	F.S/P.H.S.
14	171	428	115	289	0.15	0.08	0.21	0.10	0.21	0.12	0.17	0.10	0.18	0.10	F.S./R.H.S.

Figure 6: Risks for the different estimators.

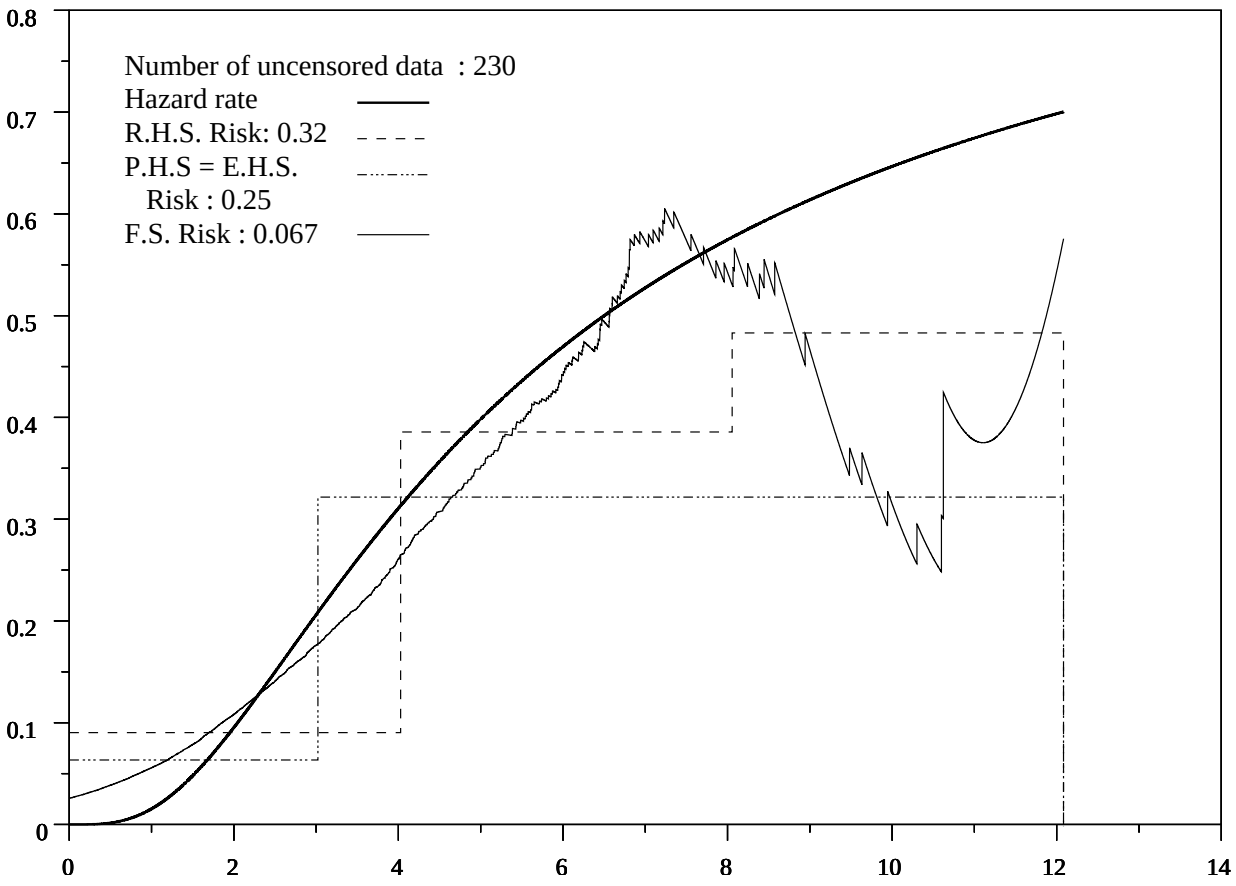

Figure 7: Estimation for a Gamma distribution. (M.C.S.=F.S)

does not achieve the minimal risk, its risk is always of the same order as the minimum of the risks.

4.3 Comparison with other existing results

In this paragraph, we want to compare our estimators to the one given by A. Antoniadis, G. Grégoire and G. Nason in [3]. Their estimator is a wavelet estimator and they choose the coefficients to keep by a cross-validation criterion. Therefore, their estimator has the same quality as ours: this is a completely data-driven non parametric estimator.

As their estimator is built on $[0, \tau]$ where τ is the last observation, we do the following rescaling: we divide the observations by τ to obtain a new set of observations in $[0, 1]$ and as the last point is always 1, we are always in Ω. This new set of observations has (if τ is deterministic) an intensity of the form $\bar{s}(t) = \tau s(\tau t)$. We estimate it on $[0, 1]$ by $\tilde{s}$ coming either from the R.H.S.($d = 2$), the P.H.S. ($d = 2.5$), the E.H.S. ($d = 0.4$), the F.S. ($d = 1$) or finally the M.C.S. Then the resulting estimator for s on $[0, \tau]$ is $\hat{s}(x) = \tilde{s}(x/\tau)/\tau$.

In the first set of simulations, the X_i's follow a Gamma distribution with shape parameter 5 and scale 1 and the U_i's follow an exponential distribution with mean 6. The results are displayed in Figure 7.

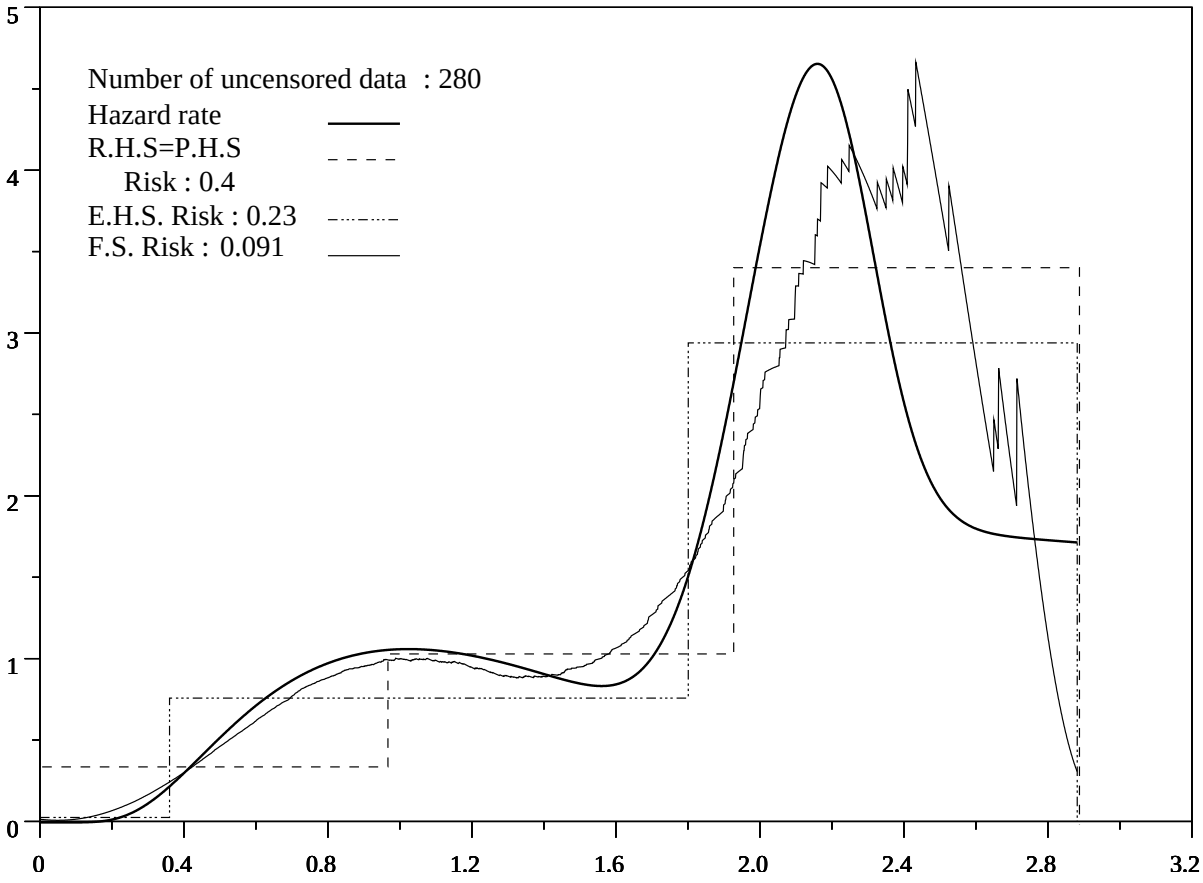

Figure 8: Estimation for a bimodal distribution. (M.C.S.=F.S)

In the second set of simulations, the X_i's have a bimodal density defined by

$$f = 0.8g + 0.2h,$$

where g is the density of $\exp(Z/2)$ with Z having a standard normal distribution and where h is the density of $0.17Z + 2$. The U_i's have an exponential distribution with mean 2.5. The results are displayed in Figure 8.

In both cases, we see that all the estimators (and especially the F.S.) are very inefficient at the end of the interval since by construction one has few observations towards the end of the interval.

We can compare our estimators with theirs by computing the same error on a lot of simulations. If one takes K regularly spaced points in $[0, \tau]$, denoted by t_k, the AMSE error is defined by:

$$\text{AMSE} = \frac{1}{K} \sum_{k=1}^{K} (\hat{s}(t_k) - s(t_k))^2.$$

The AMSE2 error is defined for the first simulation by the same kind of mean squared error but only for the t_k's less than 6. This is done in order to remove the effect of scarcity of the observations. One has $\mathbb{P}(X > 6) = 0.25$.

For the second simulations, after discussion with G. Grégoire, the AMSE2 is done for

Distributions		Gamma		Bimodal	
Number of observations		200	500	200	500
	K				
AMSE	16	0.0644	0.0554	3.050	3.090
	32	0.0786	0.0554	4.060	1.820
	64	0.112	0.0995	2.080	1.970
AMSE2	16	0.0058	0.0059	0.182	0.295
	32	0.0026	0.0021	0.152	0.066
	64	0.0025	0.0016	0.048	0.032

Figure 9: Results of A. Antoniadis, G. Grégoire and G. Nason. (see [3, Table 2])

the t_k's less than 2. One has here that $\mathbb{P}(X > 2) = 0.16$. (There is a small misprint in their article where 2.5 is written instead of 2 which is inadequate since $\mathbb{P}(X > 2.5) = 0.02$.)

All the errors are computed over 200 simulations.

We recall the results of [3] in Figure 9. As their procedure of estimation depend on the t_k's, there are three possible choices for the partitions.

We give AMSE, AMSE2 and the risk for our estimators in Figure 10. As our procedures do not depend on the choice of the t_k's, we find the same order of magnitude for each possibility. The results presented here are given with 64 points regularly spaced.

We see that the histograms strategy are better than theirs on the whole intervals in both cases. This is due to the fact that histograms do not oscillate as the end of the interval where there are less and less observations : they are more robust. On the other hand, histograms give larger result on the shorter intervals, because they are less smooth than the F.S. strategy. The F.S. strategy which is also, in this case, the one chosen by M.C.S., gives results of the same order as the result of [3]. The F.S. is better for the whole interval (especially for the Bimodal hazard rate), but is worse for AMSE2 with 200 observations. However the same phenomenon appear : AMSE2 is really much smaller than AMSE in every case.

Conclusion

In conclusion, it seems that the methods introduced in Section 2 and 3 can really be used in practice, they give results of the same order as other estimators and even better ones if we want to estimate the hazard rate as far as possible (i.e. until the last observation). The F.S. for which we are not able to prove minimax results in the general case, seems to work quite well and gives smoother results than the histograms strategies. The M.C.S. which assume that the penalized criterion is close to the risk up to a constant, allows us to take almost the best estimator among an heterogeneous family of estimators (R.H.S., P.H.S., E.H.S. and F.S) and seems to be more robust than each individual strategy.

	Distributions	Gamma		Bimodal	
Number of observations		200	500	200	500
R.H.S.	AMSE	0.0333	0.0376	0.894	0.789
	AMSE2	0.0086	0.0048	0.255	0.152
	Risk	0.278	0.179	0.559	0.321
P.H.S.	AMSE	0.0275	0.0224	1.107	0.862
	AMSE2	0.0069	0.0054	0.265	0.142
	Risk	0.246	0.190	0.617	0.338
E.H.S.	AMSE	0.0431	0.0315	1.384	0.832
	AMSE2	0.0123	0.0059	0.363	0.175
	Risk	0.397	0.243	0.865	0.415
F.S.	AMSE	0.055	0.0579	1.259	1.122
	AMSE2	0.0032	0.0012	0.150	0.051
	Risk	0.138	0.0817	0.426	0.183
M.C.S.	AMSE	0.055	0.0579	1.289	1.103
	AMSE2	0.0032	0.0012	0.160	0.051
	Risk	0.138	0.0817	0.437	0.185

Figure 10: Results of the penalized projection estimators.

5 Proofs of the main results

5.1 Proof of Theorem 1

Proof. Let d be a real number larger than 1 and let ε be a positive continuous function of d which we will choose later. Let $\Omega(d)$ be the following event:

$$\left\{ \forall I \in \Gamma, |(N_I/n) - a_I| \leq \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})}\beta_I, \right.$$
$$\left. |(N_I/n) - \alpha_I| \leq \frac{\varepsilon}{1+\varepsilon}\alpha_I, |b_I - \beta_I| \leq \frac{\varepsilon}{1+\varepsilon}\beta_I \right\}.$$

Let us bound the probability of $\Omega(d)^c$.

$$\mathbb{P}(\Omega(d)^c) \leq \sum_{I\in\Gamma} \left[\mathbb{P}\left(|(N_I/n) - a_I| \geq \frac{2\varepsilon}{(K+R)(1+\varepsilon^{-1})}\beta_I \right) + \right.$$
$$\left. + \mathbb{P}\left(|(N_I/n) - \alpha_I| \geq \frac{\varepsilon}{1+\varepsilon}\alpha_I \right) + \mathbb{P}\left(|b_I - \beta_I| \geq \frac{\varepsilon}{1+\varepsilon}\beta_I \right) \right].$$

For each of these quantities one can use Bernstein's inequality, using the individual counting processes. All the quantities are sum of n independent and centered quantities. For the first probability, we have to deal with the sum of the $(1/n)\int_0^1 \mathbb{1}_I dN^i - Y^i s dt$'s which are random variables with variance $(1/n^2)\alpha_I$. For the second probability, we have to deal with the sum of the $(1/n)\int_0^1 \mathbb{1}_I dN^i - \mathbb{E}(Y^i)s dt$'s which are random variables with variance less than $(1/n^2)\alpha_I$. Each is bounded by $M = K+R$ divided by n. For the third probability, we have to deal with the sum of the $(1/n)\int_0^1 \mathbb{1}_I(Y^i - \mathbb{E}(Y_i))dt$'s which are random variables

bounded by $1/n$ with variance bounded by $(1/n^2)\beta_I$. Hence we get

$$\mathbb{P}(\Omega(d)^c) \leq 2\sum_{I\in\Gamma}\left[\exp(-n\beta_I h(\varepsilon, M, R_\Gamma)) + \exp(-n\alpha_I h'(\varepsilon, K, M)) + \exp(-n\beta_I h''(\varepsilon))\right],$$

where h, h' and h'' are positive continuous functions. Finally we get for some positive continuous function f:

$$\mathbb{P}(\Omega(d)^c) \leq 6\frac{n}{\ln^2 n}\exp(-(\ln n)^2 f(\varepsilon, \rho, \mu, K, R))$$

which is, for fixed $\eta > 0$ less, than some $C''(d, \rho, \mu, K, \|s\|_\infty)/n^\eta$.

Let us now look at $\Omega(d)$. Let m be some fixed partition in $\mathcal{M}_A$. We know that by construction

$$\gamma_A(\tilde{s}) + \mathrm{pen}(\hat{m}) \leq \gamma_A(\hat{s}_m) + \mathrm{pen}(m) \leq \gamma_A(s'_m) + \mathrm{pen}(m).$$

For any g in $\mathbb{L}^2([0,1], dt)$, let

$$\nu_n(g) = \int_0^1 g(t)\frac{dN_t - Y_t s(t)dt}{n}.$$

Using the fact that $\gamma_A(g) = \|s-g\|^2_{\mathrm{rand}} - \|s\|^2_{\mathrm{rand}} - 2\nu_n(g)$, we obtain:

$$\|s-\tilde{s}\|^2_{\mathrm{rand}} \leq \|s-s'_m\|^2_{\mathrm{rand}} + 2\nu_n(\tilde{s}-s'_m) - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m).$$

Now, for a partition m', we denote by $m\cup m'$ the partition built on the union of sets of points which are used to construct m and m'. We denote by $\chi_{\mathcal{T}}$ the square root of $\chi^2_{\mathcal{T}}$ defined in (2.4) for a set of intervals J.

Then for all $m' \in \mathcal{M}_A$, one has

$$\sup_{f\in S'_m + S'_{m'}} (\nu_n(f)/\|f\|_{\mathrm{rand}}) \leq \sup_{f\in S_{m\cup m'}} (\nu_n(f)/\|f\|_{\mathrm{rand}}) = \chi_{m\cup m'}.$$

Hence,

$$\begin{aligned} 2\nu_n(\tilde{s}-s'_m) &\leq 2\|\tilde{s}-s'_m\|_{\mathrm{rand}}\chi_{m\cup\hat{m}} \\ &\leq \frac{2}{\varepsilon}\|s-s'_m\|^2_{\mathrm{rand}} + \frac{2}{2+\varepsilon}\|s-\tilde{s}\|^2_{\mathrm{rand}} + (1+\varepsilon)\chi_{m\cup\hat{m}}, \end{aligned}$$

using twice the fact that for all a, b, θ positive numbers,

$$2ab \leq \theta a^2 + b^2/\theta.$$

Then we obtain

$$\frac{\varepsilon}{2+\varepsilon}\|s-\tilde{s}\|^2_{\mathrm{rand}} \leq \left(1+\frac{2}{\varepsilon}\right)\|s-s'_m\|^2_{\mathrm{rand}} + (1+\varepsilon)\chi^2_{m\cup\hat{m}} - \mathrm{pen}(\hat{m}) + \mathrm{pen}(m). \qquad (5.1)$$

In order to control $\chi^2_{m\cup\hat{m}}$, we have to control all the $\chi^2_{m\cup m'}$ for m' in $\mathcal{M}_A$. First we bound $\chi^2_{m\cup m'}$ by $Z^2_{m\cup m'}V_\Gamma$ since $S_{m\cup m'} \subset S_\Gamma$. We control all the $Z^2_{m\cup m'}$'s using Proposition 2 with an upper bound on $R_{m\cup m'}$ that we denote R_Γ (this is an upper bound by additivity).

As we are on $\Omega(d)$, by additivity we are on $\Omega_{m\cup m'}(\varepsilon)$ defined in Proposition 2, and we can write that for all $x_{m'}$ positive, with probability larger than $1-\exp(-x_{m'})$,

$$Z_{m\cup m'} \leq (1+\varepsilon)\left(\sqrt{\sum_{I\in m\cup m'}\frac{\alpha_I}{n\beta_I}}+\sqrt{\frac{2R_\Gamma x_{m'}}{n}}\right).$$

We choose $x_{m'} = L_{m'}|m'|+\xi$. With probability larger than $1-\Sigma e^{-\xi}$, we control all the $Z_{m\cup m'}$ and also $Z_{m\cup\hat{m}}$. After some easy computations, we get on $\Omega(d)$ with probability larger than $1-\Sigma e^{-\xi}$:

$$Z^2_{m\cup\hat{m}} \leq (1+\varepsilon)^3 R_\Gamma\frac{|\hat{m}|}{n}(1+\sqrt{2L_m})^2+(1+\varepsilon)^3(1+\varepsilon^{-1})R_\Gamma\frac{|m|}{n}+(1+\varepsilon)^2(1+\varepsilon^{-1})^2\frac{2R_\Gamma\xi}{n}.$$

We now remark that we have built $\Omega(d)$ in such a way that on $\Omega(d)$,

$$V_\Gamma \leq (1+\varepsilon) \text{ and } R_\Gamma \leq (1+2\varepsilon)\tilde{R}_\Gamma.$$

Taking ε such that $(1+\varepsilon)^5(1+2\varepsilon)=d$, fixes ε and finishes the proof. ■

5.2 Proof of Corollary 1

Proof. Let us return to the proof of Theorem 1. One has

$$\|s-\tilde{s}\|^2_{\text{det}} = \|s-s^{\text{det}}_\Gamma\|^2_{\text{det}}+\|s^{\text{det}}_\Gamma-\tilde{s}\|^2_{\text{det}}.$$

On $\Omega(d)$, the random norm and the deterministic norms are equivalent for functions in S_Γ. Thus one has:

$$\|s^{\text{det}}_\Gamma-\tilde{s}\|^2_{\text{det}} \leq (1+\varepsilon)\|s^{\text{det}}_\Gamma-\tilde{s}\|^2_{\text{rand}}.$$

Then on $\Omega(d)$, we get

$$\|s-\tilde{s}\|^2_{\text{det}} \leq \|s-s^{\text{det}}_\Gamma\|^2_{\text{det}}+2(1+\varepsilon)\|s-s^{\text{det}}_\Gamma\|^2_{\text{rand}}+2(1+\varepsilon)\|s-\tilde{s}\|^2_{\text{rand}}.$$

We apply Theorem 1 to the last term and we integrate in ξ on $\Omega(d)$. We obtain after some computations

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\text{det}}\mathbb{1}_{\Omega(d)}) \leq (3+2\varepsilon)\|s-s^{\text{det}}_\Gamma\|^2_{\text{det}}+ C(d)\mathbb{E}\left(\inf_{m\in\mathcal{M}_A}\left\{\|s-s'_m\|^2_{\text{rand}}+\frac{|m|L_m}{n}R_\Gamma\right\}\right)+C'(d)R_\Gamma\frac{\Sigma}{n}.$$

Using (2.3) and exchanging the expectations and the infimum, there exist D and D' positive continuous functions such that

$$\mathbb{E}(\|s-\tilde{s}\|^2_{\text{det}}\mathbb{1}_{\Omega(d)}) \leq D(d)\left(\inf_{m\in\mathcal{M}_A}\left\{\mathbb{E}(\|s-s^{\text{det}}_m\|^2_{\text{det}})+\frac{|m|L_m}{n}R_\Gamma\right\}\right)+\frac{D'(d,\Sigma,R)}{n}.$$

On $\Omega(d)^c$, we use the fact that $\|s-\tilde{s}\|_\infty$ is bounded by $R+Kn^2$ and also the upper bound on $\mathbb{P}(\Omega(d)^c)$ given by Theorem 1 with $\eta=3$, to obtain the result. ■

5.3 Proof of Proposition 3

Proof. Let ψ be a positive function on $[0,1]$ symmetric about $1/2$, belonging to $\mathcal{H}_{1,\alpha,0}$ and such that $\psi(0)=0$. Then for all positive integers D, $\psi_D(x) = LD^{-\alpha}\psi(Dx)$ belongs to $\mathcal{H}_{L,\alpha,0}$. Let us fix the regular partition Γ of $[0,1]$ with D intervals. Let m be a set of intervals of Γ and let any u_I be the left extremity of any I in Γ.

Then

$$s_m = r + \sum_{I\in m} \psi_D(x-u_I),$$

belongs to $\mathcal{H}_{L,\alpha,r}$. Let $\mathcal{C}$ be a set such that for all m, m' in $\mathcal{C}$, $|m \triangle m'| \geq \theta D$ and

$$\log|\mathcal{C}| \geq \sigma D,$$

for θ and σ absolute constants. Such a set exists by application of a combinatorial Lemma (see [4, Lemma 8 p400]). Let $\mathcal{A} = \{s_m, m \in \mathcal{C}\}$.
Clearly, one has that

$$R(\mathcal{H}_{L,\alpha,r}) \geq \frac{1}{4} \inf_{\hat{s}\in\mathcal{A}} \sup_{s\in\mathcal{A}} \mathbb{E}(\|s-\hat{s}\|^2_{\det}).$$

But for all $m \neq m'$ in $\mathcal{C}$,

$$\begin{aligned}
\|s_m - s_{m'}\|^2_{\det} &= \int_0^1 \sum_{I\in m\triangle m'} \psi_D(t-a_I)^2 \mathbb{E}(Y_t^1)dt \\
&\geq \mu|m\triangle m'| \int_0^1 \psi_D(t)^2 dt \\
&\geq \mu\theta L^2 D^{-2\alpha} P,
\end{aligned}$$

where $P = \int_0^1 \psi^2$ depends only on α. Hence,

$$\begin{aligned}
R(\mathcal{H}_{L,\alpha,r}) &\geq \frac{1}{4}\mu\theta P L^2 D^{-2\alpha} \inf_{\hat{s}\in\mathcal{A}} \sup_{s\in\mathcal{A}} \mathbb{P}_s(\hat{s}\neq s) \\
&\geq \frac{1}{4}\mu\theta P L^2 D^{-2\alpha} \inf_{\hat{s}\in\mathcal{A}} (1 - \inf_{s\in\mathcal{A}} \mathbb{P}_s(\hat{s}=s)).
\end{aligned}$$

We next use a new version of Fano's lemma due to L. Birgé [5]: the infimum of the probabilities on the above right hand side is bounded by an absolute constant α' if the Kullback-Leibler distance is bounded by $\alpha'\log|\mathcal{C}|$. By the combinatorial lemma previously used, the Kullback-Leibler distance is sufficient to bound it by $\alpha'\sigma D$. But by taking the expectation of the classical formula for log-likelihood for counting processes, one has (see [1]):

$$\begin{aligned}
\forall m' \neq m \in \mathcal{C}, \quad K(\mathbb{P}_{s_{m'}}, \mathbb{P}_{s_m}) &= \int s_{m'}\phi(\log\frac{s_m}{s_{m'}})\mathbb{E}_{s_m}(Y_t)dt \\
&\leq \int \frac{(s_m - s_{m'})^2}{s_m}(x)\mathbb{E}_{s_m}(Y_t)dt \\
&\leq \frac{1}{r} nMPL^2 D^{-2\alpha}.
\end{aligned}$$

Finally, one fixes D such that

$$\frac{1}{r} nMPL^2 D^{-2\alpha} \simeq \alpha'\sigma D.$$

This leads to the result. ■

5.4 Proof of Theorem 2

Proof. Let d be a positive real larger than 1 and let ε be a positive continuous function of d that we will choose later. On Ω, we can perform the same computations as in the histogram case to obtain:

$$\|s - \tilde{s}\|^2_{\text{rand}} \leq \|s - s'_m\|^2_{\text{rand}} + 2\nu_A(\tilde{s} - s'_m) - \text{pen}(\hat{m}) + \text{pen}(m)$$

where for all g in $\mathbb{L}^2([0,1], dt)$,

$$\nu_A(g) = \int_0^1 g(t)\frac{dM_t}{A}.$$

On Ω, one can see that

$$\chi(m \cup \hat{m})_1 = \sup\{\nu_A(f)/f \in S_{m\cup\hat{m}}, \|f\|_{\text{rand}} = 1\}.$$

Therefore using the same method as for the histograms, we obtain

$$\frac{\varepsilon}{2+\varepsilon}\|s - \tilde{s}\|^2_{\text{rand}} \leq \left(1 + \frac{2}{\varepsilon}\right)\|s - s'_m\|^2_{\text{rand}} + (1+\varepsilon)\chi(m \cup \hat{m})^2_1 - \text{pen}(\hat{m}) + \text{pen}(m). \quad (5.2)$$

Moreover one has $\chi(m \cup \hat{m})^2_1 \leq \chi(m)^2_1 + \chi(\hat{m})^2_1$.

But for all m' in $\mathcal{M}_A$, we can apply the exponential formula derived in [20, Proposition 6]: for all $x_{m'}$ positive with probability larger than $1 - 2\exp(-x_{m'})$

$$\chi(m')_1 \leq \sqrt{C(m')_1} + 3\sqrt{2v_{m'}x_{m'}} + b_{m'}x_{m'},$$

where

- $v_{m'}$ is a deterministic bound on $C(m')_1$,
- $b^2_{m'}$ is a deterministic bound on $\sum_{\lambda\in m'} \varphi^2_\lambda/(Y'A^2)$.

Under the assumptions of the theorem, we obtain that for all $x_{m'} > 0$, with probability larger than $1 - 2\exp(-x_{m'})$,

$$\chi(m')_1 \leq \sqrt{\frac{|m'|}{A}}\left[\sqrt{R} + 3\sqrt{2Rx_{m'}} + \sqrt{\frac{\Phi}{c}}x_{m'}\right].$$

Let $\xi > 0$ and let $x_{m'} = L_{m'} + \xi/|m'|$. Then we can bound $\chi(m')^2_1$ by

$$(1+\varepsilon)\frac{|m'|}{A}\left(\sqrt{R}(1 + 3\sqrt{2L_{m'}}) + \sqrt{\frac{\Phi}{c}}L_{m'}\right) + (1+\varepsilon^{-1})(1+\varepsilon)\frac{18\xi}{A} + (1+\varepsilon^{-1})^2\frac{\xi^2}{A}.$$

Taking $d = (1+\varepsilon)^2$, which fixes ε, it follows that with probability larger than

$$1 - 2\sum_{m'\in\mathcal{M}_A}\exp\left(-L_{m'} - \frac{\xi}{|m'|}\right),$$

one has that

$$\frac{\varepsilon}{2+\varepsilon}\|s - \tilde{s}\|^2_n \leq \left(1 + \frac{2}{\varepsilon}\right)\|s - s'_m\|^2_n + 2\text{pen}(m) + 2\left[(1+\varepsilon^{-1})(1+\varepsilon)\frac{18\xi}{A} + (1+\varepsilon^{-1})^2\frac{\xi^2}{A}\right].$$

It remains to integrate in ξ. We finally obtain the result by a change of variables and the Beppo-Levy Theorem. ■

References

[1] P.K. Andersen, O. Borgan, R. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer Series in Statistics, 1993.

[2] A. Antoniadis. A penalty method for nonparametric estimation of the intensity function of a counting process. *Ann. Inst. Statist. Math.*, 41(4):781–807, 1989.

[3] A. Antoniadis, G. Grégoire, and G Nason. Density and hazard rate estimation for right-censored data by using wavelet methods. *J. R. Statist. Soc.*, 61(Part 1):63–84, 1999.

[4] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *P.T.R.F.*, 113:301–413, 1999.

[5] L. Birgé. A new look at an old result : Fano's Lemma. Prépublication 632, Universités de Paris VI et Paris VII, 2001.

[6] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[7] L. Birgé and P. Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 2001.

[8] G. Castellan and F. Letué. Estimation of the cox regression function via model selection. in F. Letué's PhD Thesis, UPS, 2001.

[9] L. Cavalier and J.-Y. Koo. Poisson intensity estimation for tomographic data using a wavelet shrinkage approach. September 2000, manuscript.

[10] S. Döhler and L. Rüschendorf. Adaptive estimation of hazard functions. 2000.

[11] G. Grégoire. Lest squares cross-validation for counting process intensities. *Scand. J. of Statist.*, 1993.

[12] W.-C. Kim and J.-Y. Koo. Inhomogeneous Poisson intensity via information projections onto wavelets subspaces. May 9, 2000, manuscript.

[13] C. Kooperberg, C.J. Stone, and Y.K. Truong. The L_2 rate of convergence for hazard regression. *Scand. J. Statist.*, 22:143–157, 1995.

[14] C.L. Mallows. Some comments on C_p. *Technometrics 15, 661-675*, 1973.

[15] P. Massart. Some exponential bounds for the khi-square statistics with applications. To appear.

[16] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Prob.*, 18(3):1269–1283, 1990.

[17] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Ann. Proba.*, 2000.

[18] H. Ramlau-Hansen. Smoothing counting process intensity by means of kernel functions. *Ann. Stat.*, 11:453–466, 1983.

[19] P. Reynaud-Bouret. Concentration inequalities for inhomogeneous Poisson processes and adaptive estimation of the intensity. Technical Report 18, Université de Paris-Sud, 2001.

[20] P. Reynaud-Bouret. Exponential inequalities for counting processes. Technical Report 1102-001, Georgia Institute of Technology, School of Mathematics, ftp://ftp.math.gatech.edu/pub/preprints/available_preprints_som.html, 2002.

[21] E. Rio. Une inégalité de Bennett pour les maxima de processus empiriques. Technical report, Université de Versailles-St Quentin en Yvelynes, 2001.

[22] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Stat.*, pages 65–78, 1982. Theory Appl. 9.

[23] S. van de Geer. Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Stat.*, 23(5):1779–1801, 1995.